Summer 8-15-2019

From Single Cells to Human Disease: High-resolution Phenotyping of Male Infertility Models Using Single-cell RNA Sequencing

Min Jung
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Human & Statistical Genetics

Dissertation Examination Committee:
Donald F. Conrad, Chair
Heather A. Lawson, Co-Chair
Joseph C. Corbo
Christina A. Gurnett
Benjamin D. Humphreys
Kelle H. Moley
Tim B. Schedl

From Single Cells to Human Disease: High-resolution Phenotyping of Male Infertility Models Using Single-cell RNA Sequencing

by

Min Jung

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

"It takes a whole village to raise a ~~child~~ PhD student" – "modified" African proverb

My last 5 years at WashU were filled with so much awe and love. With endless love and support from the DBBS community, Conrad lab, faculty and friends, I have grown to become a better scientist and a person. I am sincerely grateful for this transformative yet wonderful journey that is about to conclude.

Of course, nothing was possible without my amazing thesis mentor, Don. I am very thankful for his unwavering support, patience and optimism. He surely fostered an exciting working environment and I enjoyed every single second of my time in his lab.

I thank my thesis committee, Dr. Heather Lawson, Dr. Christina Gurnett, Dr. Tim Schedl, Dr. Joseph Corbo, Dr. Benjamin Humphreys, and Dr. Kelley Moley for their constructive feedbacks and support that really helped to shape this thesis work.

I also want to thank previous and current Conradians: Michiel, Ni, Nic, Amy, Arthur, Jannette, Avi, Chengran, Brian, Abul, Wu-Lin, Ana, Liina, and Nicole. I would not forget all the laughter and amazing food that we shared together. I want to especially express my gratitude to couple lab members: Jannette for our random conversation on just about everything and helping me with generating Drop-seq data; Abul for the help with maintaining the mouse colony; Ana for being the best partner in crime in the lab and bringing insightful scientific discussion which allowed us to publish two co-authored manuscripts together.

Outside of the lab, I had great support from "LMN" girls: Lilian and Nicole. I am 99.99% sure that I would have quit PhD long time ago if I didn't meet them. They were always there for me, rain or shine. We shared so much tears, joy and TMI stories throughout my PhD, allowing us to grow tighter. We served as an inspiration to each other to push our limits in our research and

personal lives. I will not forget about all the practice talks, scientific discussion, fun excursions outside of lab and our Kaldi's therapy.

Lastly, I want to thank my loving parents, Brother Jeff and my furry friends, Kermit and Fozzie. My lovely parents always served as an inspiration to put my best foot forward and have compassion for people around me. I am very lucky to have Jeff as my brother, who is also pursuing a PhD in computer science here at WashU. He has been always a loyal friend who shows endless support and care for me. This dissertation would not have been possible without Jeff's support. Thank you, my furry friends, Kermit and Fozzie for greeting me with kisses and strong headbutts every time come home.

Min Jung

Washington University in St. Louis

August 2019

Cheers to coffee, baking therapy, my loving parents, Brother Jeff, and friends

“The future is amazing!” - D.F.C.

ABSTRACT OF THE DISSERATION

From Single Cells to Human Disease: High-resolution Phenotyping of Male Infertility Models Using Single-cell RNA Sequencing

by

Min Jung

Doctor of Philosophy in Biology and Biomedical Sciences

Human & Statistical Genetics

Washington University in St. Louis, 2019

Professor Donald F. Conrad, Chair

Professor Heather A. Lawson Co-Chair

Male infertility is a complex disease that can result in significant emotional distress and treatment costs. Globally, male infertility affects 7% of males, and while its incidence is rising, its etiology remains elusive. In order to improve patient care, it is critical to identify the nature of spermatogenic failure in as many men as possible. The extensive cellular heterogeneity of testis has limited the application of bulk expression measurements to capture crucial information to dissect molecular mechanisms of defects in the infertile patients. Thus, the application of single-cell RNA-sequencing on male germ cells provides an amazing new set of scientific opportunities for research in male reproductive biology and translational medicine. We developed a single-cell framework that utilizes high-throughput single-cell RNA-sequencing from normal and disease models to elucidate normal spermatogenesis, and to dissect spermatogenic defects in male infertility models. As part of our single-cell framework, we first a developed fast, efficient yet high-throughput single-cell isolation method that can be easily applied to different mammalian species. Using Drop-seq, we generated a 57,600-cell dataset from testes of wild-type mice and

mice with gonadal defects due to disruption of the genes, *Mlh3, Hormad1, Cul4a or CNP*. For analyzing this novel data, we introduce a model-based factor analysis method, Sparse Decomposition of Arrays (SDA), to jointly analyze mutant and wildtype cells and decompose our data into latent factors ("components") that represent genes that co-vary across subsets of cells. Our single-cell framework identified novel cell-type specific markers, co-regulated gene modules and mutant-specific pathological processes. It also led us to identify a rare population of macrophages within the seminiferous tubules of *Mlh3-/-*and *Hormad1-/-*models, an area typically associated with immune privilege. These results demonstrate the potential of our single-cell framework for expanding the ability to dissect pathophysiology in tissues with extensive cellular heterogeneity and decrypt the spermatogenic failure in more patients.

Chapter 1: Introduction

1.1 Spermatogenesis

Spermatogenesis is a male version of gametogenesis that generates functional male gametes required for fertility. Spermatogenesis is initiated when spermatogonial stem cells mitotically divide into two types of spermatogonia, type A and B. The type A spermatogonium has three possible fates: it can differentiate into another type A spermatogonium, replenishing the stem cell pool; it can undergo apoptosis; or it can differentiate into the committed spermatogonium, also known as type B spermatogonium (**Figure 1**). Maintenance of the spermatogonial stem cell population is crucial for ensuring continuous spermatogenesis throughout the reproductive lifespan of males. However, it is currently unclear what causes these undifferentiated spermatogonia to choose the fate toward the committed differentiation over self-renewal.

During spermatogonial division and subsequent stages of spermatogenesis, cytokinesis is not complete and male germ cells remain connected to one another by cytoplasmic bridges[1](**Figure 1**). Ions and molecules will readily pass through these intercellular bridges, ensuring synchronous development of germ cell cohort[2]. Type B spermatogonia mitotically divide into primary spermatocytes and these primary spermatocytes will enter the first round of meiotic division to generate secondary spermatocytes (**Figure 1**). Then secondary spermatocytes will undergo another round of meiotic division to generate haploid male germ cell type called spermatids (**Figure 1**). Since the second round of meiotic division happens very rapidly, these cell types are harder to identify in histological sections[3]. The resulting spermatids are still connected by cytoplasmic bridges and are functionally diploid as transcript molecules will be shared through the bridges[2].

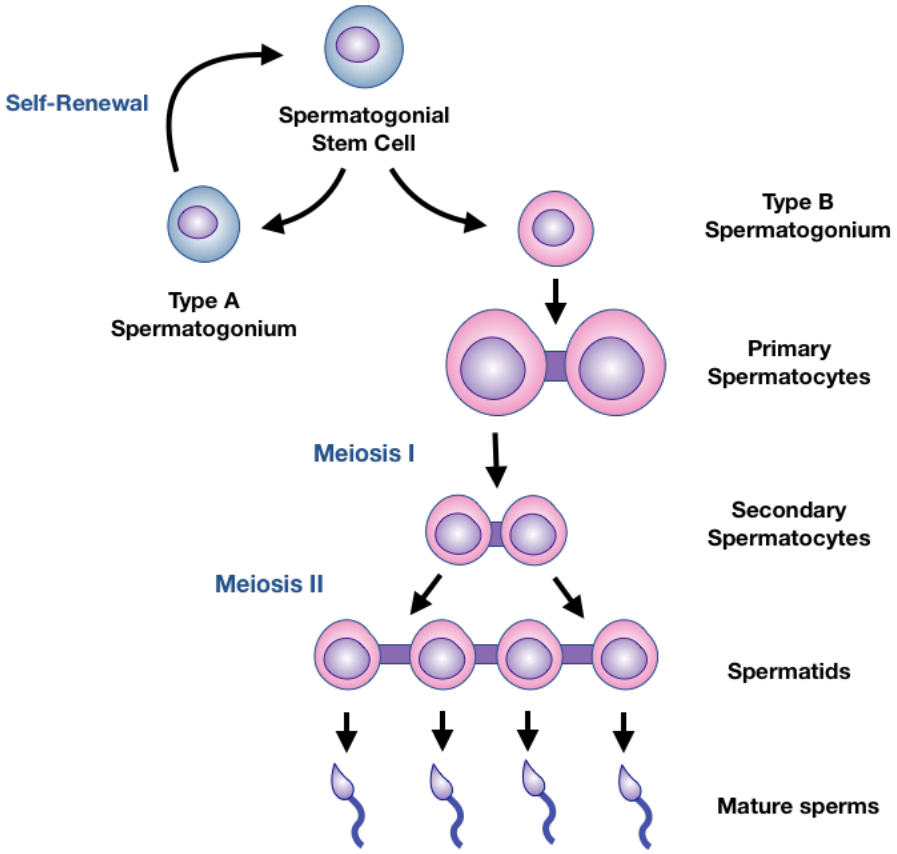


Figure 1: Overview of spermatogenesis.

The resulting round and unflagellated spermatids that are morphologically distinct from mature sperms will go through a series of maturation steps which is known as spermiogenesis. Spermiogenesis prepares the spermatids to gain crucial functional features including fertilization and motility. The first steps of spermiogenesis consist of forming acrosomal vesicles from the Golgi apparatus on one end of the nucleus and centrioles migrate to the other end of the nucleus. The nucleus of differentiating spermatid starts to condense as the acrosome cap forms and flagellum develops from migrated centrioles. The last stage of spermiogenesis begins when the elongating spermatid's remaining cytoplasm gets jettisoned as the nucleus continues to flatten and condense. The excess cytoplasm, also known as the residual body, is phagocytosed by surrounding Sertoli cells. Finally, the spermiogenesis completes as the mature spermatids or

spermatozoa are released by Sertoli cells into the lumen of the testes. As spermatogenesis progresses, male germ cells move farther away from the basement membrane of the seminiferous tubules and closer to the lumen of the tubule. Therefore, distinct male germ cell types are observed in particular layer of the tubule.

The process of spermatogenesis requires highly specific interplay between different somatic and germ cell populations. Somatic cells including Sertoli cells and Leydig cells, play crucial roles in every germ cell differentiation stage. The Sertoli cells, which are in close contact with spermatogenic cells, nourish and protect the developing germ cells to successfully mature into sperms. They are also crucial for maintaining Leydig cell population and for the support of normal peritubular myoid cell function[4]. Leydig cells produce a group of androgen hormones which includes testosterone and these hormones are required for maintaining homeostasis of spermatogenesis.[5] Myoid cells are contractile cell types that are involved in the transport of spermatozoa and testicular fluid in the tubule.[6] A number of studies have alluded that these cells communicate Sertoli cells through growth factors, affect Sertoli cell functions and achieve normal spermatogenesis. Therefore, it is crucial to maintain the optimal biological conditions for the process to complete correctly.

The study of testis biology is fascinating as it features a number of unique features that may provide insights into studying stem cell biology, developmental gene regulation, epigenetics, adaptive evolution and fertility[7–10]. First, testis transcriptome has the largest number of tissue specific genes, which is over twice as many as the 2ndranked tissue, the cerebral cortex – with which the testis shares an unusual similarity. [11–13]. A deeper understanding of the transcriptional program of spermatogenesis has potential applications in contraception [14], *in vitro* sperm production for research and the treatment of infertility [15], and the diagnosis of infertility, among

others. Testis contains the only cells in the male body with sex chromosome inactivation [16]. Meiotic sex inactivation is a protective mechanism that sequesters sex chromosomes which do not align during meiotic cell division in germ cell development. In addition, meiotic cells undergo programmed double strand break formation, homologous chromosome pairing, and recombination. As discussed above, cells undergoing meiosis share transcripts through cytoplasmic bridges [2]. Finally, male germ cells feature one of the most dramatic chromatin remodeling processes in our body, when the majority of histones are stripped away during spermiogenesis and replaced with small, highly basic proteins known as protamines [17]. It is noted that the transcription of the protamine gene is seen in the early haploid cells (spermatids), although translation is delayed for several days.[18] The tightly packed chromatin results in complete shutdown of transcription in maturing spermatids.

1.2 Male infertility

Infertility affects approximately 15% of couples globally and 45% of these infertile cases are described to be male factor. Despite its prevalence and societal importance, the underlying etiology of male infertility remains unknown in up to 40% of the cases.[19–21]Considering more than 2000 genes are involved in male gametogenesis, it is likely that the majority of idiopathic male infertility is caused by genetic mutations in the spermatogenesis candidate genes, affecting relevant physiological processes such as hormonal homeostasis, spermatogenesis, and sperm quality.[22,23] However, despite an extensive search for new genetic factors involved in male reproductive biology, no clinically relevant gene mutations have so far been identified in the past two decades due to the presence of extensive cellular heterogeneity within testis and technical limitations of using low resolution molecular technologies such as karyotyping or polymerase chain reaction (PCR) for diagnosing these patients.[22]

In clinics, male infertility diagnosis begins with evaluation of reproductive history, physical examination and semen analysis. Semen analysis measures parameters such as semen volume, sperm concentration, sperm motility, viability, and morphology.[24] However, significant overlap exists in these clinical laboratory values between normal and infertile men.[25] Thus, discovering additional novel biomarkers would be helpful to differentiate normal cohort and infertility patient cohort. In cases of men with severe spermatogenic impairment (when sperm is either extremely scarce or completely absent from the ejaculate), testicular biopsies are collected. In addition to providing a source of germ cells for assisted reproduction, the biopsy is usually histologically characterized (e.g. using Hematoxylin and Eosin staining) to provide a better sense of the nature of the gonadal defect (i.e. developmental arrest, complete lack of germ cells). The current clinical approach of dissecting gonadal dysfunction in male infertile patients is very qualitative and phenomenological, limited to describing what can be observed in histological sections. Moreover, the male reproductive medicine field doesn't use a conceptual framework for assigning cases to predefined "types" of molecular defects – impairment of the blood testis barrier, failure of spermiation, etc.

Recent advances in single-cell RNA-seq technologies could potentially circumvent the limitations described above. By capturing transcriptomic profiles at a higher resolution in individual cells, we can negate the need to select cell types to study and deconvolute the molecular heterogeneity among different testicular cell types. Using single-cell RNA-sequencing, we can generate detailed anthology of gene expression phenotype and changes underlying spermatogenic development. This framework will allow us to take a step towards identifying "failure modes" for spermatogenesis which will accelerate our understanding of the many new infertility mutations we will encounter in the era of clinical sequencing.

1.3 Single-cell RNA-seq

Single-cell RNA-seq has emerged as a revolutionary tool that allows us to solve scientific enigma that eluded our examination previously. The development of single-cell RNA-sequencing technology is relatively recent but it has become an extremely popular technique in various scientific fields. Traditional methods, such as microarray and bulk RNA-seq, were limited to provide measurements that are averaged signals from individual cells present in the population, leading to averaging artifacts.[26–28] Given the presence of great heterogeneity within the seemingly homogenous cell population, bulk measurements do not accurately reflect the composition of tissues, the dynamics of transcription, and the regulatory relationships between genes.[26,27]

The first single-cell RNA-sequencing data was published in 2009 by Tang and his colleagues.[29] In their paper, they successfully surveyed the first single-cell transcriptome by manually isolating cells in single tubes, extracting RNA and amplifying cDNA.[29] Since then, many different advancements including introduction of microfluidics, random capture methods and in-situ barcoding, have transformed the magnitude of single-cell study. Most recently, droplet-based methods (i.e. 10X Genomics, Drop-seq and In-Drops) have emerged as the popular choice of method as they allow random capturing of single cells in droplets and carry out barcoded cDNA preparation within the droplets in a massively parallel manner.[30,31] Briefly, single-cell suspension, barcoded beads with poly(T) primers in cell lysis buffer and oil are co-flowed into a microfluidics device to generate droplets where an individual cell and individual bead is encapsulated. The barcodes on the beads are unique in that it allows us to track down which cell the mRNA transcript information came from when analyzing the data.

To convert raw sequencing reads to meaningful biological interpretation, a pipeline of computational steps is performed. These steps include quantification of gene expression, quality

control, batch correction, customized downstream analyses, and visualization.[32] We quantify gene expression through aligning and mapping sequencing reads to the genome using aligners like STAR, Kallisto and Salmon.[33,34] Quality control is performed by filtering out low quality data based on the number of UMIs per cell, number of genes per cell and mapping rates. Batch effects are technical variations are introduced when samples are processed and sequencing libraries are generated.[35–37] These batch effects can interfere with data interpretation, masking true biological signals. Once data is quality controlled and batch effects are removed, one can perform customized downstream analyses which include dimensionality reduction for visualizing the data, unsupervised cell clustering based on transcriptomic similarity, differential gene expression analysis for identifying novel markers and pseudotime analysis for constructing developmental trajectories for studying cellular dynamics.[32] There are a number of available tools for performing this workflow such as Cell Ranger, Monocle and Seurat.[38,39,40]

The potential of scRNA-seq promises an exciting future with numerous biological and medical applications.[41,42] However, it also presents unprecedented technical and computational challenges, leaving a room for further improvement.[28,41,42] Preparation of good quality single-cell suspension is crucial as poor quality of input single-cell suspension can affect downstream interpretation of data significantly. In addition, it can also be hard to generate single-cell suspension for certain type of tissues (i.e. clinical samples or tissue from animals with cell types that are hard to dissociate).[43] From the computational standpoint, one of the major hurdles comes from the high dropout rate, owing to the low amount of starting material in single cells. Therefore, the scRNA-seq data is nosier and much sparser when compared to bulk RNA-seq data. Although there have been a number of computational tools developed to recover missing transcripts or impute the 'true' gene expression pattern using different algorithms (i.e MAGIC), development of

novel methods that can effectively address the noisy and sparse nature of the scRNA-seq data is in need.[44]

The following chapters demonstrate our humble attempt to address a couple of limitations mentioned above by (1) developing a single-cell isolation method that is optimized and efficient in time, labor, and quality (2) developing a single-cell transcriptome analysis framework for dissecting gonadal defects in male infertility from sparse single-cell RNA-sequencing data.

Chapter 2: A Standardized Approach for Multispecies Purification of Mammalian Male Germ Cells

Preface:

Portions of this chapter have been reproduced and adapted from the following published manuscripts:

Lima, A. C*., **Jung, M***., Rusch, J., Usmani, A., Lopes, A., & Conrad, D. F. (2016). Multispecies Purification of Testicular Germ Cells. *Biology of reproduction*, 95(4), 85. doi:10.1095/biolreprod.116.140566

&

Lima, A. C*., **Jung, M***., Rusch, J., Usmani, A., Lopes, A. M., & Conrad, D. F. (2017). A Standardized Approach for Multispecies Purification of Mammalian Male Germ Cells by Mechanical Tissue Dissociation and Flow Cytometry. *Journal of visualized experiments: JoVE*, (125), 55913. doi:10.3791/55913

*indicates co-first authors

2.1 Abstract

Advanced methods of cellular purification are required to apply genome technology to the study of spermatogenesis. One approach, based on flow cytometry of murine testicular cells stained with Hoechst-33342 (Ho-FACS), has been extensively optimized and currently allows the isolation of nine germ cell types. This staining technique is straightforward to implement, is highly effective at purifying specific germ cell types, and yields sufficient cell numbers for high-throughput studies. Ho-FACS is a technique that does not require species-specific markers, but whose applicability to other species is largely unexplored. We hypothesized that, because of the similar cell physiology of spermatogenesis across mammals, Ho-FACS could be used to produce highly purified subpopulations of germ cells in mammals other than mouse. To test this hypothesis, we applied Ho-FACS to four mammalian species that are widely used in testis research: *Rattus norvegicus*, *Cavia porcellus*, *Canis familiaris,* and *Sus scrofa domesticus*. We successfully isolated four germ cell populations from these species with average purity of 79% for spermatocytes, 90% for spermatids, and 66% for spermatogonia. Additionally, we compare the performance of mechanical and chemical dissociation for each species and propose an optimized gating strategy to better discriminate round and elongating spermatids in the mouse, which can potentially be applied to other species. Our work indicates that spermatogenesis may be uniquely accessible among mammalian developmental systems, as a single set of reagents may be sufficient to isolate germ cell populations from many different mammalian species, opening new avenues in the fields of development and male reproductive biology.

2.2 Introduction

Spermatogenesis is a complex developmental process in which early spermatogonial stem cells differentiate into spermatozoa in the seminiferous tubules of the testes. The study of this fascinating process has produced critical insights into stem cell biology,[8] developmental gene regulation,[7] adaptive evolution,[9] and fertility.[10] With over 30 different distinct cell types in the vertebrate testis, there is exceptional diversity in the expression profiles of cells within a single individual, which can become confounding when studying expression differences among individuals or developmental stages.[45] This has compelled researchers to develop methods for effective male germ cell enrichment and isolation, such as StaPut velocity sedimentation, elutriation, magnetic-activated cell sorting (MACS), whole testis collection during the first wave of spermatogenesis, and fluorescence-activated cell sorting (FACS) with Hoechst-33342 (Hoechst).

StaPut and elutriation are fairly efficient techniques that allow separation of different germ cells based on their size and density. When applied to mice, StaPut yields about 10^8 cells/population from 22 testes with 90% purity, whereas approximately 10^7 cells/population can be obtained by elutriation of two testes with 80%–95% purity rate.[46,47] In both methods, the fractionation step that collects purified cells from different bovine serum albumin or Percoll gradients is labor intensive (3–4 h) and requires proficiency from practice as well as specific equipment. Also, both techniques are unsuitable for detailed molecular studies during meiosis as they can separate only one type of meiotic cell subpopulation at a time and fail to yield sufficient purity.[48,49] MACS, which separates desired germ cell populations by conjugating the germ cells with a known surface marker antibody primed with magnetic beads, may circumvent this issue

by performing purification in parallel with population-specific antibodies. However, only spermatogonia (Spg) and spermatids (Spd) are proven to have established surface markers for successful enrichment.[46] Furthermore, antibody-assisted purification has limitations in that it is necessary to develop species-specific reagents for each marker, and antibody-assisted purification typically does have the sensitivity to discriminate between cells at slightly different stages of a quantitative developmental process. Collecting mouse testis samples at specific days postpartum, timed for the first appearance of different germ cell types during the first wave of spermatogenesis, is also used to enrich specific germ cell populations.[50] Given that the testis size is very small at those time points, and that samples comprise a mixture of all testicular cells, this approach is experimentally challenging and fails to detect intrinsic biological variations among individual cells. Importantly, evidence from different studies suggests that the first wave is regulated differently from adult spermatogenesis.[50–52]

FACS of Hoechst-stained (Ho-FACS) murine male germ cells can discriminate nine germ cell types.[48,53–56] Hoechst is a vital dye that binds preferentially to poly(d[AT]) sequences in the minor groove of DNA, with secondary binding taking place at higher ratios. These two DNA binding sites show varying binding energies and consequent spectrum shifts in relation to chromatin amount and structure.[57,58] It has been proposed that this spectral shift could be used to discriminate between cells with similar DNA content but different chromatin properties.[59–61] Indeed, Ho-FACS of male germ cells exhibits a pattern that reflects changes in DNA content (blue fluorescence) and chromatin structure (red fluorescence) throughout spermatogenesis. In fact, red fluorescence shifts resulting from progressive chromatin decondensation during meiotic prophase allow the resolution of different meiotic subpopulations.[53,55,62] Spermatogonial stem cells are an exception and represent a side population because of BCRP1-dependent dye efflux,

which is switched off after the spermatogonial stages.[53] Therefore, measuring Hoechst intensity as a function of blue and red fluorescence is representative of three cellular properties: ploidy, chromatin structure/accessibility, and dye efflux caused by ABC transporter activity.[53,55,62,63] With over 95% purity of sorted populations[55] and an average of 10^7 cells/population from two testes in less than 2 h, this technique has proven highly efficient and less labor intensive. Although the actual FACS session requires specialized sorting equipment (ultraviolet [UV] laser) and a skilled operator, many research facilities provide cell-sorting services.

Recently, there is a growing interest in applying genomic technology to germ cells, especially in an evolutionary context.[64,65] In that sense, purified cells can be used for numerous applications ranging from studying gene regulation to nucleosome mapping, epigenetics, development in germ cells, and many more.[66–68] To unravel the complexity of germ cells at a genomic level, researchers need an efficient and high-throughput purification technique that can be applied easily to many species. Given that Ho-FACS is not based on a species-specific molecular signature (e.g., an antigen), and that the cellular machinery for spermatogenesis is similar across all vertebrates, we hypothesized that separation of different germ cell types by Ho-FACS could be applied to other species. To test this hypothesis, we applied Ho-FACS to four species that are highly valued by the testis research community: *Rattus norvegicus* (rat), *Cavia porcellus* (guinea pig), *Canis familiaris* (dog), and *Sus scrofa domesticus* (miniature pig; hereafter mini pig).

Our results provide detailed descriptions on how Ho-FACS performs with an optimized gating strategy that includes a cell viability gate with propidium iodide (PI) staining and a DNA content gate at cell enrichment for four primary types of germ cells in each of the four species that we investigate. Each of our target spermatogenic germ cell types could be distinguished by

Ho-FACS of the diploid (2C) mammalian species. We also demonstrate the use of a mechanical testis dissociation protocol in comparison to species-specific conditions for enzymatic dissociation and present an optimized FACS gating strategy based on cell shape, size, and complexity to distinguish elongating Spd (eSpd) and round Spd (rSpd) in mouse. Collectively, we offer the first proof of principle that flow cytometry can be applied transversally across mammalian species to isolate Hoechst-stained male germ cells in different developmental stages.

2.3 Results

2.3.1 Efficiency of Tissue Dissociation Protocol Is Crucial for Cell Sorting with Hoechst Staining

We isolated two testes from each animal in the study: 10 mice, 4 rats, 3 dogs, 2 guinea pigs, and 1 mini pig. In order to confirm a normal adult testis phenotype of the collected specimens, we performed HE staining of tissue sections from one testis, and then submitted the other testis for FACS. A microscopy analysis of the HE slides shows the expected tissue architecture and organization of normal adult male testes and highlights some general differences across species (**Figure. 1**). Mammals have a tubular testicular arrangement, with spermatogenesis progressing from the periphery towards the lumen, and show interspecific variability of germ cell morphology.

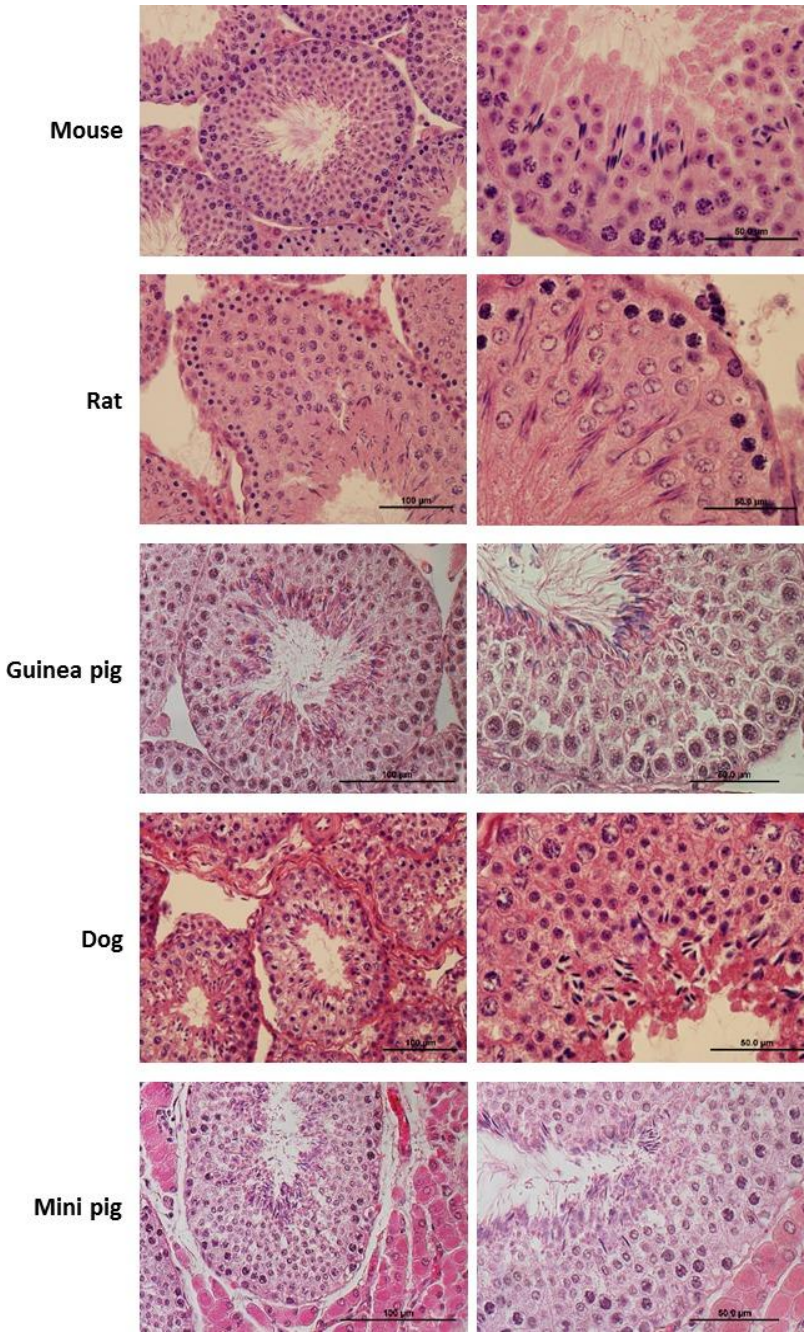


Figure 1: Histology of testicular tissues from four mammalian species.
HE staining of testicular cross sections of collected specimens. For each animal studied in the paper (n = 10 mice, n = 3 dogs, n = 4 rats, n = 2 guinea pigs, n = 1 mini pig) we processed testis fragments for histology and for FACS. Here we present representative HE staining from each species. In each subject, histological examination of testicular cross sections shows the presence of all germ cell types in different developmental stages at lower (left panel) or higher (right panel) magnification, confirming that the specimens were sexually mature and presented a normal testicular phenotype.

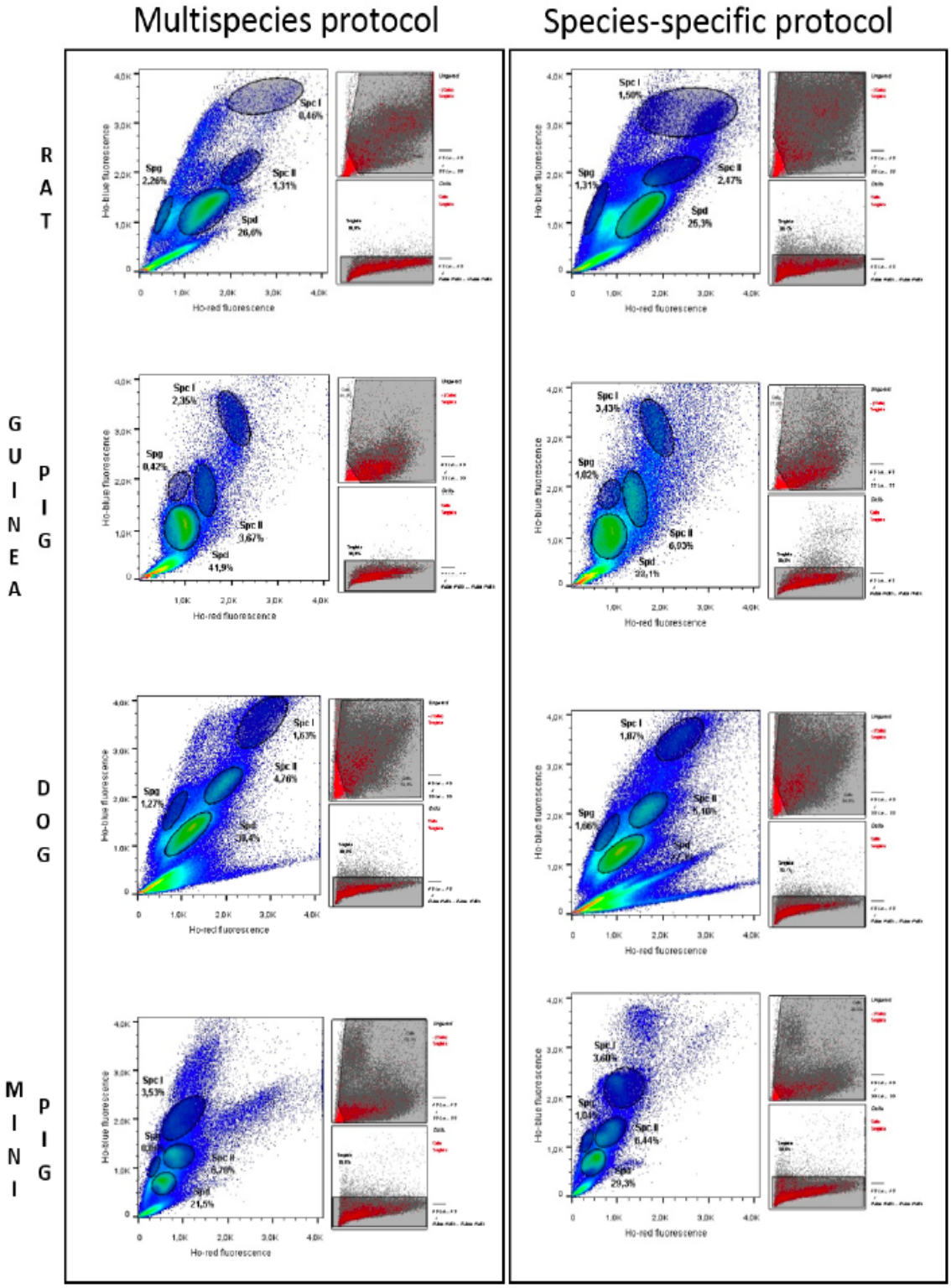


Figure 2: Ho-FACS plots of cell suspensions obtained using enzymatic dissociation protocols.

Sample preparation has direct implications in the success and results obtained by flow cytometry. In this figure, plots reflect measurements of Hoechst fluorescence of stained germ cells isolated from testes of rat, guinea pig, dog and mini pig,using an enzymatic dissociation multispecies protocol, optimized for mouse, or species-specific enzymatic dissociation protocols (See Methods and Supplementary data). In general, adjusting the enzymatic dissociation according to the species being process resulted in a higher abundance of male germ cells for each population. Percentages indicate the proportion of cells within the gates in relation to the total number of live cells. Smaller windows show the parent gates leading to the plots generated as a function of Ho-blue and Ho-red fluorescence intensities. Spg: spermatogonia; Spc I: primary spermatocytes; Spc II: secondary spermatocytes; Spd: spermatids. Plots were obtained using FlowJo® software v10 (Tree Star Inc.).

The success of cell-sorting protocols depends on the quality of inputted single-cell suspensions and is therefore directly affected by the efficiency of tissue dissociation. Here, we evaluated the use of different protocols for testis dissociation, enzymatic and mechanical, in each of the species studied. We applied an enzymatic dissociation protocol optimized for mouse testis to all species, referred to as the multispecies protocol, and defined species-specific protocols by adjusting incubation temperatures and times and/or trypsin concentrations and/or introducing the use of hyaluronidase to improve digestion of connective tissues (**Chapter 2.5 Materials and Methods**). To control for technical and biological variables, the experiments were performed simultaneously in tissue sections of the same testis for both biological replicates, except for mini pig. Overall, species-specific enzymatic dissociation protocols performed better as evaluated by the separation of distinct clusters obtained on the FACS profiles (**Figure 2**). The main goal of a tissue dissociation protocol is to reduce the amount of manipulation and time while retaining the viability of the dissociated cells. We therefore tested the applicability of mechanical testicular dissociation in different species using a method originally described for rodents.[69] To evaluate the performance of our approach, we estimated the mean percentage of cells that passed the gates during FACS and compared these values with the ones obtained for mouse. The proportion of cells from total counts (**Figure 3**) is indicative of the sample quality and a reflection of the

efficiency of tissue dissociation protocols. The average percentage of cells passing the debris filter was comparable between species, with mechanical dissociation performing similarly or better when compared to enzymatic dissociation with species-specific protocol. Notably, the percentages estimated for all species are directly influenced by the high stringency of the debris filter applied during FACS. These results suggest that cell sorting with Hoechst staining seems very sensitive to sample quality, validating our approach of designing species-specific enzymatic dissociation protocols that are effective in generating good single-cell suspensions. More importantly, mechanical dissociation with Medimachine provides a standard method for testicular dissociation that reduces the processing time and is applicable to different mammalian species.

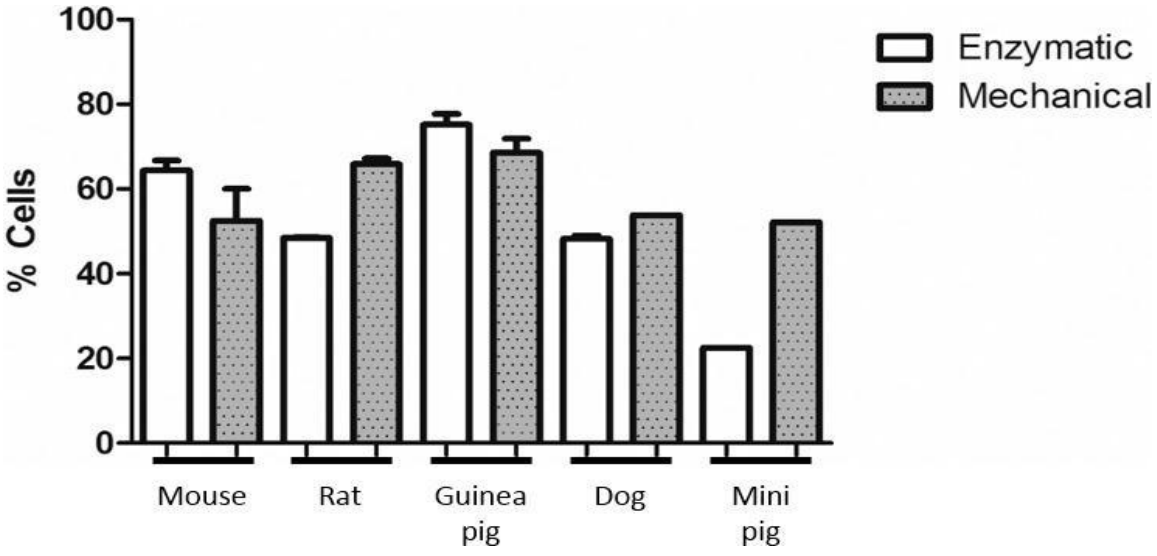


Figure 3: Evaluation of testis dissociation protocols by flow cytometry.
In order to evaluate the efficiency of mechanical and enzymatic dissociation protocols in different species we estimated the percentage of live cells from total number of cells. The quality of mechanical single-cell suspensions was comparable to the species-specific enzymatic dissociation protocol, indicating that this method can be used to quickly obtain single cell suspensions of different mammalian species. Data sets were obtained from variable numbers of experimental replicates (mouse, 10; dog, 3; rat, 4; guinea pig, 2; and mini pig, 1) using FlowJo software v10 (Tree Star Inc.). Histograms were generated with GraphPad Prism (version 5.02 for

Windows, GraphPad Software, www.graphpad.com), plotting the calculated mean values with standard deviation.

2.3.2. Male Germ Cell Types of Different Mammalian Species Can Be Discriminated by Ho-FACS

We first isolated four different populations (Spg, primary Spc [Spc I], secondary Spc [Spc II], and Spd) from dog and rat testicular cell suspensions obtained from species-specific enzymatic dissociation (**Figure 2**). The gates for sorting were defined based on the cluster of cells observed and taking into account the expected location of the subpopulations in terms of Hoechst red and blue fluorescence: 1) Spg (side population), low Hoechst blue and red fluorescence; 2) Spc I (4C euchromatin to heterochromatin), high Hoechst blue and a wide range of low to high Hoechst red fluorescence; 3) Spc II (2C euchromatin to heterochromatin), intermediate Hoechst blue and a wide range of low to high Hoechst red fluorescence; 4) Spd (1C compacted chromatin with structural variations resulting from histone to protamine transition), low Hoechst blue and a narrow range of Hoechst red. Moreover, it appears that the chromatin of the dog germ cells is generally more compact throughout spermatogenesis as the clusters of cells show a trend of lower red Hoechst fluorescence. To identify the germ cell types and quantify the purity of the sorted subpopulations, we performed a microscopy evaluation of cell morphology and chromatin structure based on Hoechst fluorescence (**Chapter 2.5 Materials and Methods**).

Immunofluorescence in tissue sections with Hoechst was used as reference for the pattern of Hoechst staining in different germ cells (**Figure 4A**). Spg are small, round cells with distinct pericentric heterochromatin. Spc are larger granulated cells, with Spc II populations defined by the detection of binucleated cells or cells in diakinesis. Spd are small 1C cells that can be round or elongated in shape. Despite the similar size, rSpd can be clearly distinguished from Spg by the

presence of localized chromocenters (**Figure 4B**). Purity was estimated based on this analysis, indicating 74%–85% purity of specific cell types passing through each gate (**Table 1**) except for the dog Spg population (46%), because of the close proximity of the eSpd and Spg populations in fluorescence space (**Figure 2**). For the Spc I gates, most contamination was with preleptotene Spc. In the rat, this could have resulted from the absence of clearly defined premeiotic and meiotic Spc subpopulations during FACS.

Species	Spg	Spc I	Spc II	Spd
Rat	81%	74%	79%	85%
Dog	46%	81%	81%	82%

Table 1. Purity of cell populations isolated by species-specific enzymatic dissociation of rat and dog testes.

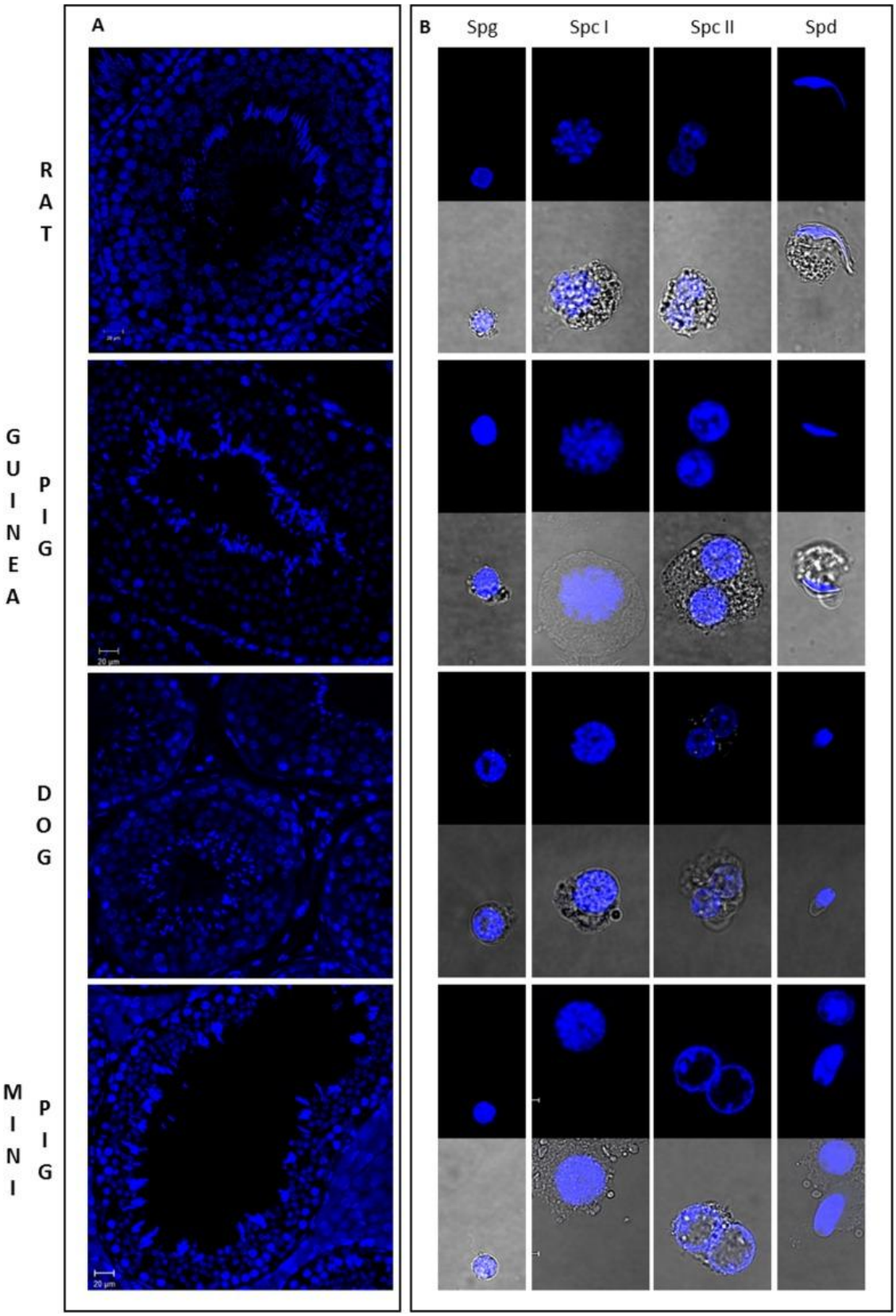

A
B
Spg
Spc I
Spc II
Spd
RAT
GUINEA PIG
DOG
MINI PIG
20 µm
20 µm

Figure 4: Microscopic evaluation of germ cell populations isolated from mammalian testes by Ho-FACS.
Immunostained cross sections of rat, guinea pig, dog, and mini pig testes (**A**) were used as reference for the classification of isolated germ cells in respect to chromatin structure marked with Hoechst (blue). Morphological evaluation of chromatin structure was performed based on cell shape and size, allowing the identification of different germ cell types (**B**). Spg are small round cells with compact heterochromatin whereas Spc I and Spc II show larger and more complex cells with more diffuse chromatin (Spc I) and/or binucleated cells (Spc II). Spd gates comprise cells in different states of spermiogenesis, ranging from rSpd to completely elongated Spd. Panels indicate the designated FACS gate. Slides were visualized in a confocal microscope. For each isolated population, Hoechst fluorescence of sorted cells was visualized after FACS and images were collected under a ×63 magnification lens, with (lower panel) or without (upper panel) white light transmission.

Then, to ensure the viability of the cells being sorted and to refine the purity of the populations obtained, we applied a gating strategy similar to that described for the mouse in Gaysinskaya et al.[55] (**Figure 5**). This strategy includes a viability gate based on PI staining as well as a DNA content gate where a histogram obtained based on Hoechst-blue fluorescence shows peaks representative of 1C, 2C, and 4C cells. Figure 6 shows the cytograms, as a function of Hoechst-blue and red fluorescence, generated during Ho-FACS of single-cell suspensions obtained from testicular tissue of all the different species by mechanical dissociation. Although we see some expected interspecific variation in the pattern of the FACS profiles, we can clearly distinguish at least four subpopulations of male germ cells for all species. The different cell populations were sorted by applying the gating strategy described in **Figure 5** and purities were estimated (**Table 2**) by a similar morphological analysis as described above (**Figure 4**). Looking at the relative frequency of cells passing through each gate (**Table 2**), a similar higher frequency of Spd was detected for all species; however, the abundance of other germ cell types varied among species. These observations were expected and presumably reflect interspecific differences in testis composition and the technical challenge of making standardized settings for subpopulation gating. Interestingly, although this gating strategy generally improved the purity

of germ cell populations isolated from the rat and dog (**Table 1 and Table 2**), the guinea pig and mini pig Spg populations showed contamination with other cell types and an overall lower level of purity. Altogether, our results suggest that Ho-FACS, combined with PI staining, of testis single-cell suspensions can be used to isolate germ cells from rat, guinea pig, dog, and mini pig, further strengthening our hypothesis that this method can potentially be applied as a generalized procedure for isolation of germ cells in different mammalian species.

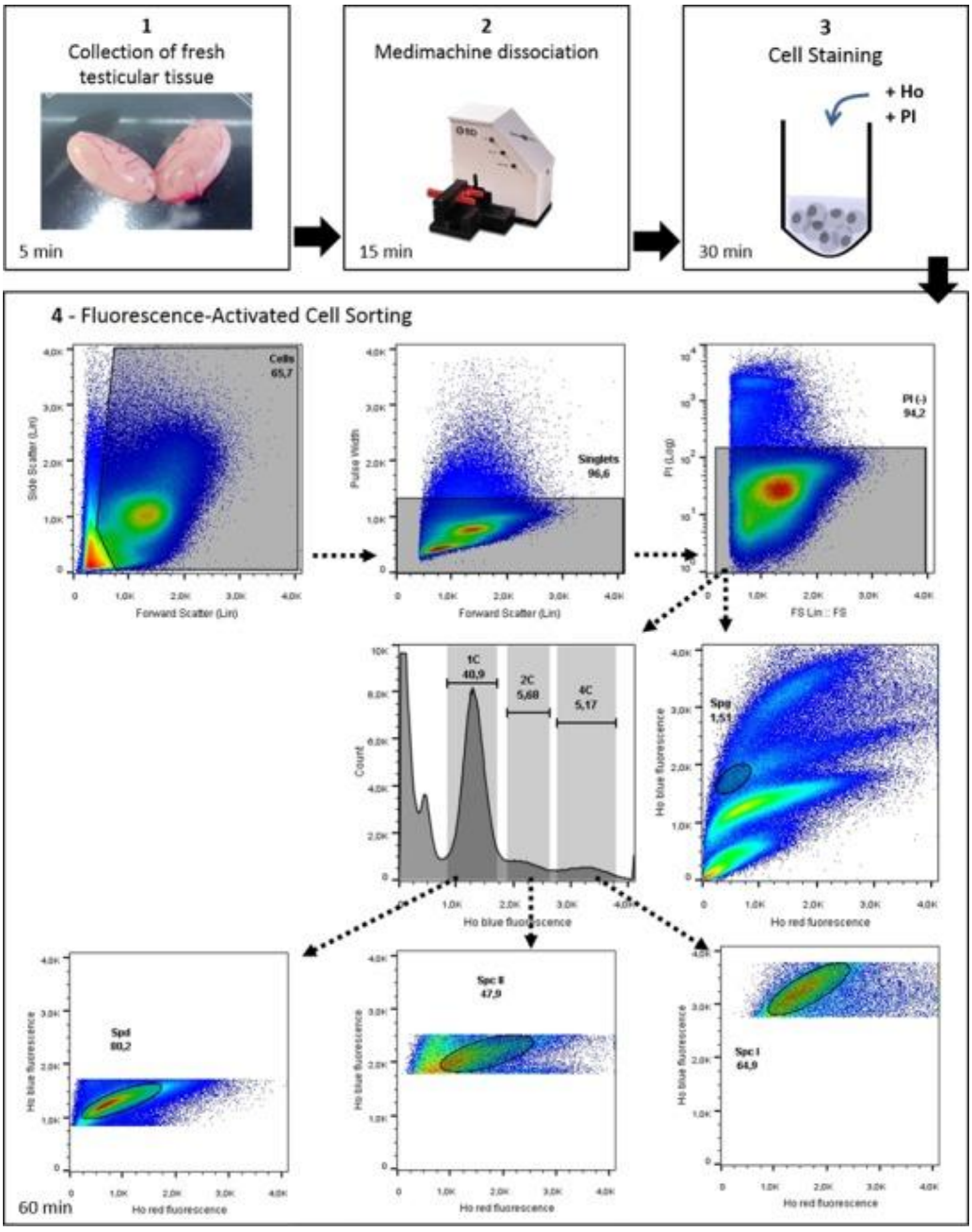


Figure 5: Workflow of Ho-FACS isolation of mammalian male germ cells.
This image summarizes the steps for germ cell isolation of mammalian germ cells, represented here by the application of this protocol to rat testis. The BD Medimachine system is used for mechanical tissue dissociation. FACS is performed in a Beckman Coulter MoFlo Legacy cell sorter (see Materials and Methods) and plots generated using FlowJo software v10 (Tree Star Inc.). Ho: Hoechst.

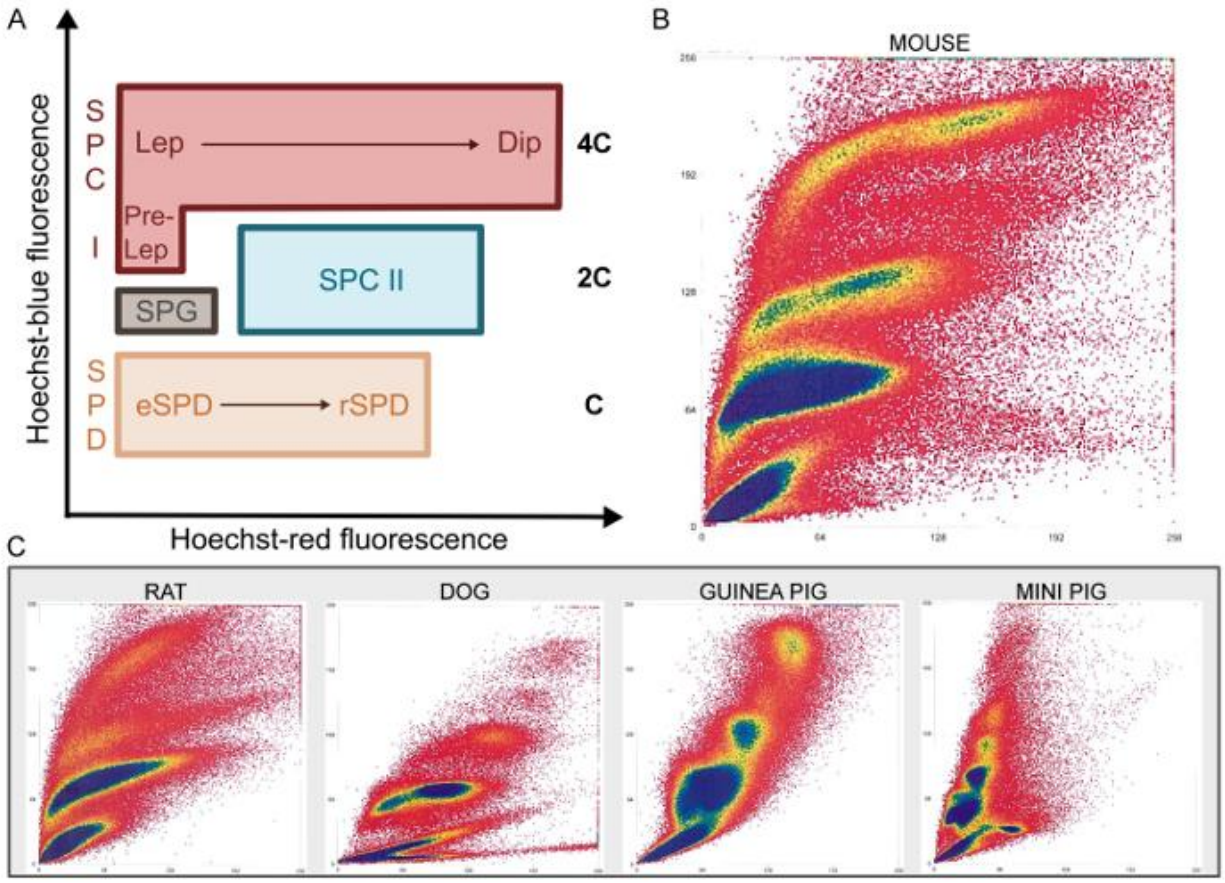


Figure 6: Interspecific comparison of Ho-FACS plots of testicular single-cell suspensions. (A) Live cells are then analyzed based on Hoechst fluorescence: blue is proportional to DNA content and increasing red fluorescence reflects less condensed chromatin and structural variations. As such, male germ cells of different stages are expected to cluster in specific regions of cytograms plotting the function of blue/red Hoechst fluorescence. (B-C) Plots represent the ratio of blue and red Hoechst fluorescence obtained by flow cytometry after testis mechanical dissociation and staining of germ cells of the mouse, rat, guinea pig, dog, and mini pig. Gating (round circles) was defined based on observed cell clusters and expected location of populations in relation to Hoechst fluorescence. A minimum of four populations were identified and sorted for all species. Ho: Hoechst.

			Cells in DNA content gate (%)			Purity of sorted germ cell populations (%)					
Species	Singlets (%)	Live cells (%)	1C	2C	4C	Spg	Spc I	Spc II	Spd	rSpd	eSpd
Mouse	98.4	92.5	34.7	3.9	3.72	74	82	87.5	95.2	95*	92*
Rat	95.7	93.8	37.7	5.3	5.4	83	81	82	87	–	–
Guinea pig	96.2	92.1	39.3	7.6	5.8	48	68.7	85	87	–	–
Dog	97.9	86.5	16.4	3	0.5	78	–	87	–	91	81
Mini pig	95.9	93.2	26.9	6.4	3.5	49	52	82	92	–	–

* Obtained by enzymatic dissociation and gated based on FSC and SSC parameters (see text).

Table 2. Statistics of Ho-FACS of male germ cell suspensions obtained by mechanical dissociation.

2.3.3. rSpd and eSpd Can Be Separated by Ho-FACS Based on Cell Shape and Size

During Ho-FACS of the different mammalian species, it was notable that while Hoechst-red and blue fluorescence alone could discriminate rSpd and eSpd populations in the dog sample, it was insufficient to further refine this population in the remaining species (**Figure 6 and Table 2**). Given that rSpd and eSpd are molecularly very distinct in terms of transcription activity as well as the differentiation occurring in the latter during spermiogenesis, we sought to evaluate a different strategy to isolate different mouse spermatid subpopulations by FACS. It has been previously suggested that rSpd and eSpd could be gated based on high and low FSC, respectively.[53] Interestingly, we observed that gating based on the FSC parameter alone introduced some contamination in the sorted populations. Microscopy quantification of purity of sorted populations based on cell morphology and Hoechst fluorescence revealed enrichment of ~62% for eSpd and 84% for rSpd (**Figure 7**). Gating for events with low FSC and high versus low SSC, we increased purity levels to 92% for eSpd and 86% for rSpd (**Figure 7**). Finally, we observed that the lowest levels of contamination could be obtained by the combination of FSC and SSC gating followed by Hoechst red/blue fluorescence. It seems that eSpd can be isolated

gating for low FSC and SSC with 83%–92% enrichment range, whereas rSpd appear to have higher FSC and SSC values and can be separated with 86%–95% accuracy (**Figure 8 and Table 2**). Importantly, this gating strategy is based on cell size, shape, and complexity and thus potentially applicable to Ho-FACS of any species undergoing spermiogenesis during gamete development.

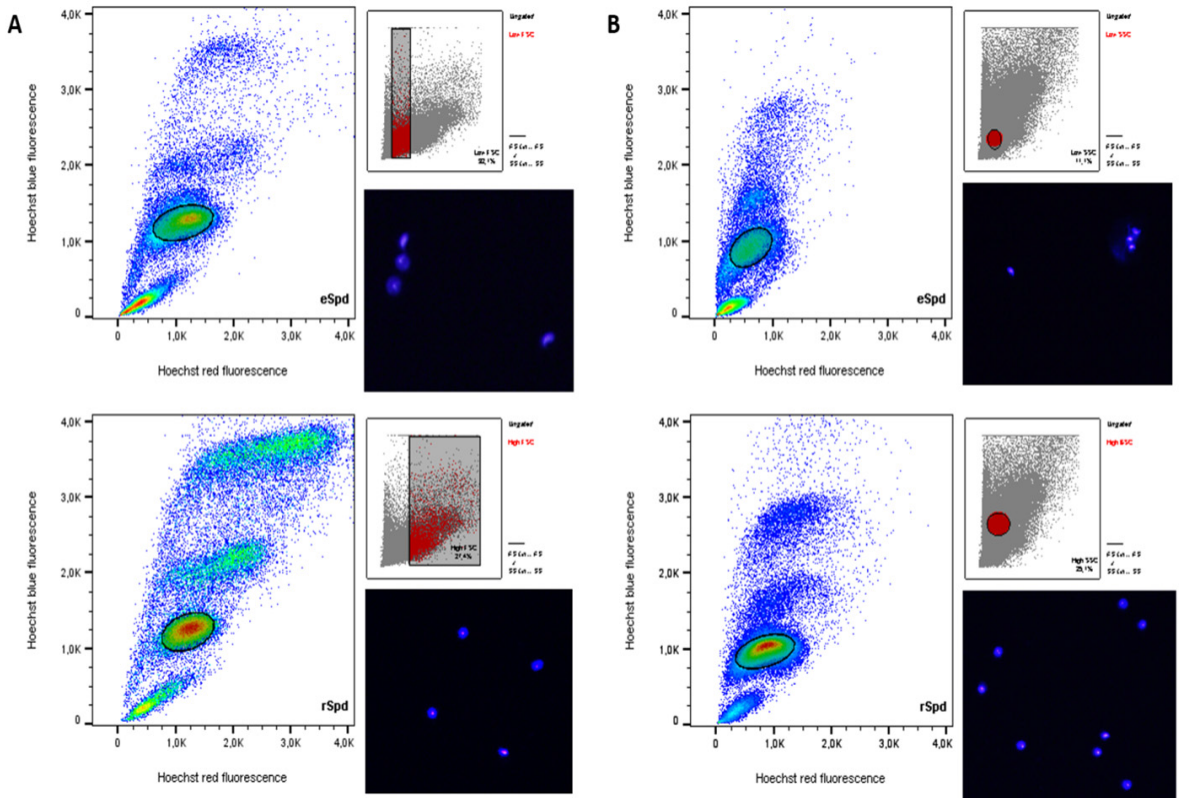


Figure 7: Optimization of a gating strategy to isolate round and elongating spermatids.
In order to discriminate between round (rSpd) and elongating spermatids (eSpd) we defined the parent gates (circles and squares with cells labeled red) to reflect differences in cell shape (A) or complexity (B). Gates for sorting were then defined by the expected pattern of Hoechst blue/red fluorescence for spermatids. Cell populations gated for high or low FSC were enriched 62% for rSpd and 84% for eSpd respectively (A). Within a range of low FSC, gating for higher or lower SSC increased the enrichment to 86% and 92% of rSpd and eSpd, respectively, in the sorted population (B). Morphology of sorted cells was evaluated based on Hoechst fluorescence and images were acquired by light microscopy with a UV lamp (16X magnification lens.). Plots were generated using FlowJo software v10 (Tree Star Inc.).

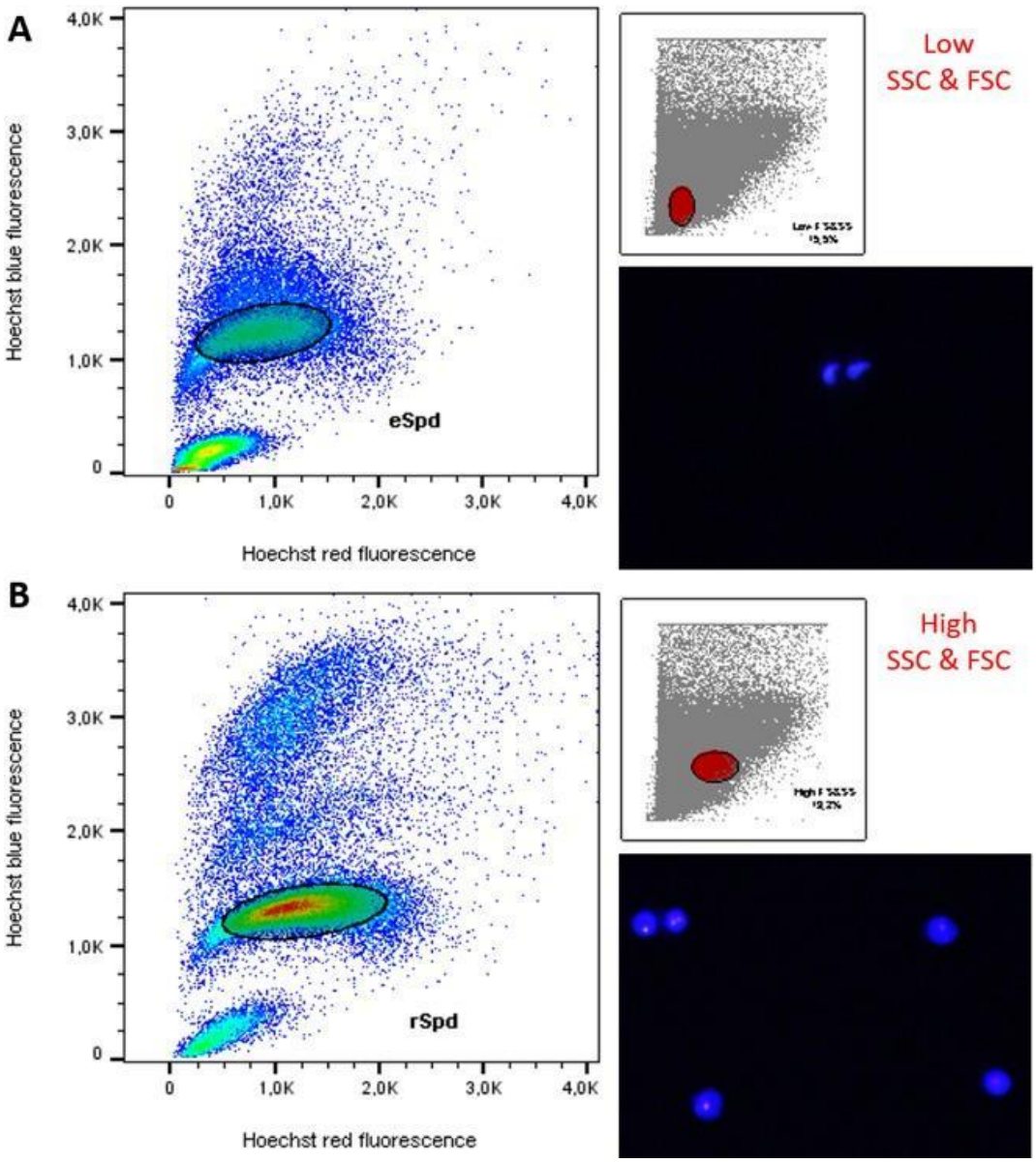


Figure 8: Gating strategy to discriminate rSpd and eSpd.
Cell shape and complexity influence the ratio of FSC and SSC parameters measured by flow cytometry. The smaller windows in both images show the parent gate (red full circle) based on FSC and SSC. Gated cells then clustered as functions of Hoechst blue/red fluorescence with the pattern expected for 1C cells with condensed chromatin, defining the gates for sorting. Morphology of sorted cells was evaluated microscopically based on Hoechst fluorescence and confirms the enrichment of the expected cell types in each population. Therefore, eSpd are smaller and less complex, showing lower ratios of FSC and SSC (A), whereas rSpd present higher FSC and SSC (B). Cell images were obtained by light microscopy with a UV lamp (×16 magnification lens). Plots were generated using FlowJo software v10 (Tree Star Inc.).

2.4 Discussion

One of the major challenges in male reproductive biology has been to design a method to differentiate and isolate subtypes of developing germ cells with a high percentage of recovery and low contamination with other cell types. Since the first reports over a decade ago, flow cytometry of Hoechst-stained murine male germ cells has been slowly revisited and optimized to isolate premeiotic (Spg), meiotic (preleptotene, leptotene, zygotene, pachytene, diplotene, and Spc II) and postmeiotic (rSpd and eSpd) stages.[48,53–56,70] This technique is based on measurements of chromatin amount and structure detectable using the fluorescent Hoechst DNA dye. Flow cytometry of testicular cell suspensions from nonmouse mammalian species using different dyes, staining protocols, and flow cytometry parameters for analysis has revealed similar profiles in terms of DNA ploidy/stainability (reviewed in Geisinger and Rodriguez-Casuriaga[71]). We reasoned that, for species sharing similar chromatin dynamics (2C-4C-2C-1C) and structure throughout spermatogenesis, major populations of germ cells in different stages could be isolated by Ho-FACS. Here, we propose a workflow that (**Figure 5**) is fast and straightforward and allows the simultaneous isolation of 4 germ cell populations from fresh testicular tissue in less than 2 h. The reduced processing time is crucial to maintain cellular integrity for further downstream procedures. Moreover, its successful performance in 5 different species suggests it could be broadly applied within the mammalian clade, making it the ideal method to isolate germ cells for comparative studies of mammalian male reproductive biology.

Using three rodent species (mouse, rat, and guinea pig) and two non-rodent species (dog and mini pig), we showed that the general resolution of distinct cell populations is maintained across mammals and allows the isolation of at least four developmental stages: Spg, Spc I, Spc

II, and Spd. The purity of these subpopulations was slightly reduced when compared to previous works for the mouse,[53,55] but shows good enrichment of expected cell types (**Table 2**). It is important to highlight that higher purities of early and mid-late Spc I (93%–97%) have been described for FACS sorting of germ cells with >2 h of PI staining,[72] suggesting that longer incubation periods increase the power of discriminating germ cells based on DNA-binding dyes in flow cytometry. Moreover, this could also explain a lower percentage of Spc I cells detected in the guinea pig and rat samples. Nonetheless, a reduced processing time is crucial to preserve the physiology of ex vivo cells and, in that regard, a combination of Hoechst and PI staining for 30 min seems to be sufficient for a good enrichment of different male germ cell types (**Figure 5**). The presence of eSpd was the major source of contamination within the Spg gates and resulted from their close spatial proximity in Hoechst plots, reaching the highest values in the guinea pig and mini pig FACS. One possible way to circumvent this issue would be to stain germ cells with a membrane permeable marker for the acrosome, allowing to gate cells for the presence of this spermatid-specific structure. In fact, Spg is the most challenging population to isolate based on Hoechst staining. Spermatogonial stem cells show a unique Hoechst fluorescence pattern and represent a side population as a result of BCRP1-dependent dye efflux.[53] Other methods such as MACS using Spg-specific markers would be more suitable to isolate spermatogonial stem cells for studies focusing on this particular cell type. Nonetheless, a sample preparation method achieving an enrichment of even 50% Spg is likely to have an important impact as a useful tool in the world of germ cell genomics, especially in single-cell studies.

Future work would also include the optimization of this protocol to discriminate other cell types in different mammals. Here, we describe an optimized gating strategy based on cell size, shape, and complexity to differentiate rSpd and eSpd in the mouse (**Figure 7 and Figure**

8), suggesting that the isolation of populations enriched for these germ cells can be achieved for other mammalian species. Also, discrimination between different meiotic stages, already resolved for mouse,[53–55] would broaden the scope of the application of this technique in the field of male reproductive biology.

Overall, we provide the first evidence supporting the applicability of Ho-FACS as a transversal method to isolate male germ cells in different developmental stages across mammalian species. As a proof of principle, our work has major implications for several types of studies in developmental biology. First, it provides the tools to investigate the dynamics of germ cell development in different species individually, which would benefit research of understudied mammalian species such as domesticated animals.[73] Furthermore, using the same experimental procedure in different species reduces noise and eliminates sources of variables that often challenge comparative studies. In the "omics" era, with the growing interest in applying genome technology to address questions about epigenetics, regulation, and protein diversity throughout spermatogenesis,[51,52,64–68,74–77] this technology could be used to comprehensively tackle different aspects of germ cell development with an evolutionary perspective.

2.5 Materials and Methods

Animals

C57BL/6 male mice (Jackson Laboratory), Sprague Dawley male rats (Harlan Bioproducts), guinea pigs, and mini pigs were raised in animal facilities at Washington University in St. Louis. Dog testes were collected at Hillside Animal Hospital (St. Louis, MO) from animals scheduled for castration, and were transported to the lab on ice for immediate processing. Prior to surgery, dogs are routinely injected with lidocaine and bupivacaine to help with the recovery process. All

testis samples were obtained from sexually mature animals (mice, 8–12 wk; rats, 70 days; dogs, 12–24 mo; guinea pigs, 3 mo; and mini pigs, 6 mo) and procedures were conducted in compliance with regulations of the Animal Studies Committee at Washington University in St. Louis.

Collection and Processing of Testicular Tissue

Fresh testes from each species were decapsulated, rinsed in 1× PBS (#AM9625; Thermo Scientific), and cut to the size of mouse testis (approximately 1.5 × 0.7 cm). These tissue fragments were used without further processing for dissociation and FACS sorting or fixed for histology. For immunofluorescence, tissue was fixed in 4% paraformaldehyde (PFA; #15710; VWR) overnight at room temperature and washed with 70% ethanol at least three times. Testes sections used for hematoxylin-eosin (HE) staining were collected in modified Davidson solutions (24 h at room temperature with gentle rotation; #64133-50; Electron Microscopy Sciences), fixed in Bouin solution (24 h at room temperature with gentle rotation; #HT101128; Sigma), and washed with 70% ethanol until any remaining yellow color of Bouin fixative was completely removed.

Immunofluorescence and HE Staining

Fixed testes samples were processed in an ethanol series and embedded in paraffin and 5-μm sections were cut. Slides were deparaffinized with xylene and rehydrated to PBS through sequential ethanol washes with decreasing alcohol concentrations. Standard HE staining was performed according to HE protocol adapted from Belinda Dana (Department of Ophthalmology, Washington University in St. Louis School of Medicine) with Hematoxylin 560 (#3801570;

Surgipath) and 1% Alcoholic Eosin Y 515 (#3801615; Surgipath) for overall morphological evaluations. Immunofluorescence staining was performed after antigen retrieval (boiling in citric acid buffer for 20 min) and tissue permeabilization/blocking (0.5% Triton X-100 + 2% goat serum in 1× PBS for 1 h at room temperature). Primary (anti-P-H3[ser10]; #Ab5176; AbCam) and secondary (goat anti-rabbit ALF 633; #A21071; Life Technologies) antibodies were diluted (1:100 and 1:500 respectively) in antibody dilution buffer (1× PBS + 1% Tween 20 + 1% BSA) and incubated overnight at 4°C and 4 h at room temperature, respectively, in a humid chamber. After secondary antibody incubation, sections were stained with Hoechst (1:500; #H3570; Life Technologies), washed with 1× PBS, and mounted with ProLong Diamond Antifade Mountant (#P36961; Life Technologies). For comparative purposes with FACS-sorted germ cells, only Hoechst fluorescence is shown from these sections.

Testis Dissociation and Hoechst Staining

Two different types of testicular dissociation protocols were used in this work: enzymatic and mechanical. The latter was performed using a Medimachine system (Cat. #340588; BD Biosciences) in line to the method previously described for rodents in Rodriguez-Casuriaga et al.[29] A multispecies enzymatic dissociation protocol was designed based on the procedure described in Margolin et al.[8] for mouse, as described below, and species-specific adjustments were made in terms of enzymes used, their relative concentrations, and incubation time and temperatures. Except for mini pig, whose species-specific adjustments were made according to Park et al.,[30] enzymatic dissociation conditions were adopted following the Worthington references for reproductive tissue dissociation.[31]

Enzymatic dissociation of testicular tissue (multispecies protocol)

Solutions (fresh; prior to testes collections) were prepared as follows: collagenase type I (120U/ml; Worthington Biochemical, #LS004196) + cycloheximide (CHX; 0.1 mg/ml; Amresco #94217) in 1× Dulbecco modified Eagle medium (DMEM; #31053; Life Technologies); trypsin (50 mg/ml; #LS003708; Worthington) in 10 mM HCl; and DNase I (1 mg/ml; #10104159001; Roche) in 50% glycerol.

1. Testis enzymatic digestion: Testes (mouse) or testes fragments (rat, dog, guinea pig, and mini pig) were placed in 15-ml conical tubes containing 3 ml of DMEM/collagenase I/CHX solution and 10 µl of DNase I solution. The tube was shaken vigorously until the testicular tubules started to disperse and then agitated horizontally at a speed of 120 rpm for 15 min at 33°C. Temperature and agitation speed were the same for all subsequent incubation steps.
2. Somatic cell removal: Tubules were decanted for 1 min vertically at room temperature and the supernatant was discarded to remove somatic cells.
3. Seminiferous tubule digestion: 2.0 ml of DMEM/collagenase I/CHX, 50 µl of trypsin, and 10 µl of DNase I solutions were added and the tube was inverted several times. After a 15-min incubation period, the tubules were gently pipetted up and down for 3 min using a plastic disposable Pasteur pipet with a wide orifice. Then, 30 µl of trypsin and 10 µl of DNase I solution were added and the tube was inverted several times, followed by another 15 min digestion period.
4. Staining with Hoechst: 400 µl of fetal bovine serum (FBS; #10082139; Thermo Scientific) was added and mixed by inverting to inactive trypsin, followed by addition of 5 µl of Hoechst and 10 µl of DNase I. The cell suspension was incubated for 15 min, then

passed through two 40-µm 1× DMEM-prewetted disposable filters and stored on ice and protected from light until ready for FACS processing (not more than 30 min).

Species-specific alterations of enzymatic dissociation protocol

In order to evaluate the effect of testis dissociation protocols in FACS, tissue sections of all specimens were also dissociated using species-specific protocols based on the procedure described as multispecies protocol with the following modifications:

1. Rat and Guinea pig: Trypsin stock concentration was adjusted to 1mg/ml and Hyaluronidase (1mg/ml; Sigma-Aldrich #H6254) stock solution was added to the 1X DMEM/Collagenase I. The last incubation time in step 4 was adjusted to 20 min.
2. 2. Dog: Collagenase stock concentration was adjusted to 0.2% by dissolving 20mg Collagenase type I in 10ml 1X DMEM. The first incubation time in seminiferous tubule digestion was adjusted to 30min.
3. Mini pig: 0.1% Collagenase type 4 and 0.1% Hyaluronidase were added to 1X DMEM and trypsin concentration adjusted to 0.25%. All incubations were carried out at 37°C.

Mechanical dissociation of testicular tissue

Two to three testis fragments of ~2–3 mm^3 were placed in a Petri dish containing 1 mL of ice cold phenol red free 1x Dulbecco's Modified Eagle Medium (DMEM), cut with a scalpel, and transferred to a prewetted disposable Medicon disaggregator with 50-µm separator mesh (Cat # 340591; BD Biosciences). Tissue was processed for 5 min and resulting cell suspension was recovered from the Medicon unit with a 3-ml disposable syringe, passed through two 40-µm 1× DMEM-prewetted disposable filters, and transferred back to the Medicon unit for 5 min more of processing. The resulting single cell suspension was transferred to a 1.5-ml tube and stained with

10 µl of Hoechst and 2 µl of PI for 30 min at room temperature in the dark. Samples were then filtered again (40-µm filter) and kept on ice in the dark until FACS processing.

Fluorescence-Activated Cell Sorting

Cells were sorted and analyzed by a Beckman Coulter MoFlo Legacy cell sorter, using Summit Cell Sorting software, similarly to the descriptions in Getun et al.[8] Hoechst was excited using a UV laser and triggered for scatter by a 488-nm blue laser. To detect Hoechst's wide emission spectrum, the UV laser was paired with a 463/25-nm band-pass filter for detecting Hoechst blue and a 680-nm LP band-pass filter for Hoechst red. A 555DLP dichroic was also used to distinguish blue from red emission wavelengths. Samples were analyzed using a 70-micron nozzle and the sorting flow rate was set to 1000–2000 events/sec. A minimum of 500 000 events were detected before proceeding to gating. Two different gating strategies were used. For cell suspensions prepared by enzymatic tissue dissociation without PI staining (rat and dog; Figure 2), a sequential cell gating strategy was applied: debris was excluded based on forward scatter (FSC) versus side scatter (SSC) plot, then singlets were gated by adjusting threshold for FSC pulse width, and finally red/blue Hoechst fluorescence was used to detect different spermatogenic germ cell populations. For samples obtained by mechanical dissociation or enzymatic dissociation with PI staining (guinea pig and mini pig) two intermediate gating steps were introduced as previously described by Gaysinskaya and Bortvin[14] for the mouse: debris was excluded based on FSC versus SSC plot, then singlets were gated by adjusting threshold for FSC pulse width. Live single cells, negative for PI, were gated based on PI red fluorescence and FSC and plotted in histograms of cell counts in relation to measurements of Hoechst blue fluorescence. Three peaks with increasing Hoechst fluorescence could be detected representative

of haploid (1C), diploid (2C), and tetraploid (4C) cells. This DNA content gate was used to refine populations of spermatocytes (Spc) and Spd, which were then discriminated by finally plotting the function of Hoechst-blue and red fluorescence intensity. Spermatogonial stem cells were identified from PI-negative cells as a direct measurement of Hoechst fluorescence because they represent a side population resulting of Hoechst efflux and therefore Hoechst-blue is not representative of DNA content in these cells. Single stained cell suspensions for Hoechst and PI were used to set optimal photomultiplier tube voltages.

Each testis was processed for 45 min to 1.5 h to collect an average of 6.0 × 10^6 cells for each subpopulation. Cells were collected into 1 ml of 1× DMEM + 10% FBS in 5-ml polypropylene round-bottom tubes or 1.5-ml tubes that were precoated with FBS. To concentrate the samples and remove dead cells and cellular debris, sorted cells were pelleted by centrifugation (600 × *g* at 4°C for 10 min) and washed in 1 ml of ice-cold 1× PBS.

Microscopic Evaluation of Purified Cells

To identify the cell types gated in each FACS-sorted population, we evaluated chromatin structure and cell morphology microscopically based on Hoechst fluorescence. During the wash step after FACS, 100 µl of sorted cells was collected, fixed in 4% PFA, and stored at 4°C in the dark. Slides were mounted with 20 µl of fixed cells from each population and visualized in upright confocal or light microscopes. To quantify cell purity, images were obtained from a minimum of 5–15 random fields and/or at least 100 cells (when available) were counted to estimate contamination with other cell types. To avoid human errors, cells were counted independently by two researchers and an estimated average cell count was reported.

2.6 Acknowledgement

We thank the Hillside Animal Hospital, (St. Louis, MO) for dog testes; Jason Arand and Dr. Ted Cicero's lab at Washington University in St. Louis (WashU) for providing rat testes and Brianne Tabers for helping with the collection; Jared Hartsock and Dr. Salt's Lab at WashU for the guinea pig testes; and Dr. Michael Talcott at the Division of Comparative Medicine at WashU for the miniature pig testis. We also thank the Alvin J. Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, MO, for the use of the Siteman Flow Cytometry Core, which provided staff-operated cell-sorting service.

Chapter 3: Unified Single-cell RNA-seq Analysis of Male Infertility Models

Preface:

Excerpted from previously published manuscript:

Min Jung*, Daniel Wells*, Jannette Rusch, Suhaira Ahmed, Jonathan Marchini, Simon Myers, Donald F. Conrad. Unified single-cell analysis of testis gene regulation and pathology in 5 mouse strains. bioRxiv 393769; doi: https://doi.org/10.1101/393769

*indicates first co-authors

3.1 Abstract

To fully exploit the potential of single-cell functional genomics in the study of development and disease, robust methods are needed to simplify the analysis of data across samples, time-points and individuals. Here we introduce a model-based factor analysis method, SDA, to analyze a novel 57,600-cell dataset from the testes of wildtype mice and mice with gonadal defects due to disruption of the genes *Mlh3*, *Hormad1*, *Cul4a*or *Cnp*. By jointly analyzing mutant and wildtype cells we decomposed our data into 46 components that identify novel meiotic gene-regulatory programs, mutant-specific pathological processes, and technical effects, and provide a framework for imputation. We identify, *de novo*, DNA sequence motifs associated with individual components that define temporally varying modes of gene expression control. Analysis of SDA components also led us to identify a rare population of macrophages within the seminiferous tubules of *Mlh3-/-*and *Hormad1-/-*mice, an area typically associated with immune privilege.

3.2 Introduction

The testis is an amalgamation of somatic cells and germ cells that coordinate a complex set of cellular interactions within the gonad, and between the gonad and the rest of the organism (**Figure 1A**). The key function of the testis is to execute spermatogenesis, a developmental process that operates continually in all male adult mammals in order to generate genetically diverse gametes through recombination and independent assortment of homologous chromosomes. The mechanisms of this process are important for the evolution, fertility and speciation of all sexually reproducing organisms.

A deeper understanding of the transcriptional program of spermatogenesis has potential applications in contraception [14], *in vitro* sperm production for research and the treatment of

infertility [15], and the diagnosis of infertility, among others. Prior to the advent of single-cell genomics, studies of the highly dynamic transcriptional programs underlying sperm production were limited by the cellular complexity of the testis. It is comprised of at least 7 somatic cell types, and at least 26 morphologically distinct germ cell classes [80].

The testis has a number of unique features: its transcriptome has by far the largest number of tissue specific genes (over twice as many as the 2nd ranked tissue the cerebral cortex – with which the testis shares an unusual similarity) [11–13]; it contains the only cells in the male body with sex chromosome inactivation[16]; meiotic cells undergo programmed double strand break formation, homologous chromosome pairing, and recombination; cells undergoing meiosis share transcripts through cytoplasmic bridges [2]; and it features dramatic chromatin remodeling, when the majority of histones are stripped away during spermiogenesis and replaced with small, highly basic proteins known as protamines [17].

Use of genetic tools has enabled the dissection of the homeostatic mechanisms that regulate spermatogenesis, revealing both cell autonomous and non-autonomous mechanisms. However, most perturbations that disrupt spermatogenesis also change the cellular composition of the testis, frustrating the use of high throughput genomic technologies in the study of gonadal defects. By removing heterogeneity as a confounding factor, single cell RNA sequencing (scRNA-seq) promises to revolutionize the study of testis biology. Likewise, it will completely change the way that human testis defects are diagnosed clinically, where testis biopsy is the standard of care for severe cases of male infertility [81].

Here, we performed scRNA-seq on 57,600 cells from the mouse testis, using wild-type animals and 4 mutant lines with defects in sperm production (**Figure 1B**). We set out to develop an analysis approach that would allow us to extract mechanistic insights from joint interrogation

of these multiple mouse strains; to gain insights into spermatogenesis and its regulation, using the precise resolution of single-cell analysis; and to establish the utility of scRNA-sequencing for dissecting testis gene regulation in both normal and pathological states.

To do this, we leverage a recently developed factor analysis method, called sparse decomposition of arrays (SDA), which has not previously been applied to single-cell RNA-seq data, and demonstrate how it can be used on scRNA-seq data for cleanup and imputation, identification of co-regulated genes, and to create a dictionary of disease from a joint analysis of mutant and wildtype animals. We show that, unlike standard clustering, we are able to decompose expression patterns into temporally overlapping yet distinct components, which each possess specific functions, providing new insights relative to recent reports of scRNA-seq from mouse testis [82–86]. Moreover, as with standard scRNA-seq analysis methods, we retain the ability of other existing scRNA-seq analysis methods to order cells from early to late meiosis, and to identify distinct groups of non-meiotic cells.

3.3 Results

3.3.1. Mapping the cellular diversity of the testis with single-cell RNA-seq

To isolate individual cells for data generation, we initially tested two methods for testis dissociation: enzymatic dissociation, a slow 2-hour protocol, vs. a rapid 30-minute protocol based on mechanical disruption [87]. Single cell expression profiles from the two methods showed excellent agreement (r=0.95), with no important differences in cell quality or ascertainment (**Figure 1, Figure 1 - Figure Supplement 1**), so we applied the mechanical dissociation approach for further experiments (**Table 1**). We performed scRNA-seq to generate 25,423 cell profiles isolated from total testis dissociations of 11 wild-type animals (WT1-WT11). We compared these

to reference data for 296 spermatogonia, 199 primary spermatocytes, 398 secondary spermatocytes, and 299 spermatids, purified by FACS (**Methods**). Transcript yield per wildtype cell (**Figure 1C**) were consistent with previous studies using DropSeq on testicular cells [83] or different cell types.

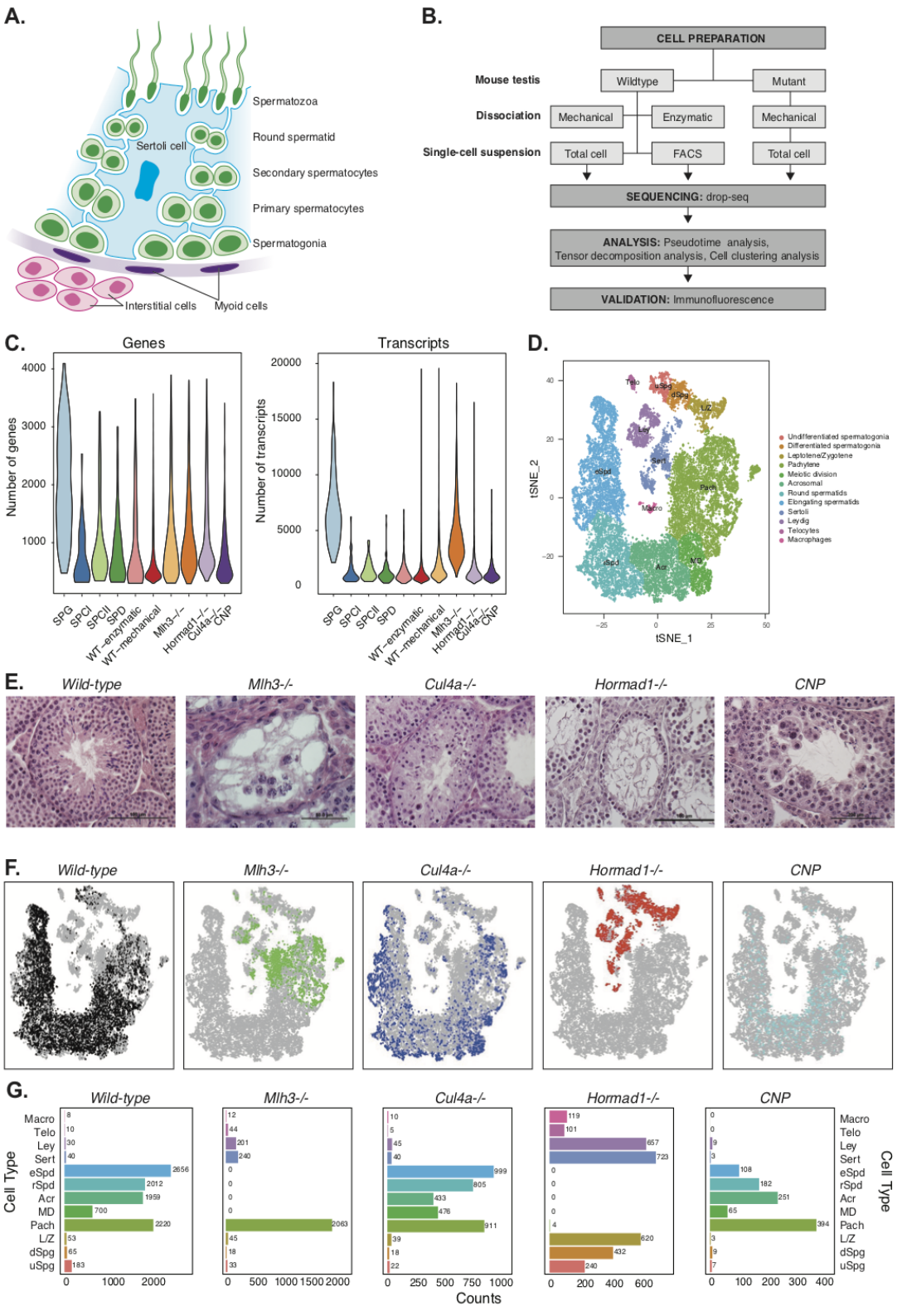

A.
Spermatozoa
Round spermatid
Secondary spermatocytes
Primary spermatocytes
Spermatogonia
Sertoli cell
Interstitial cells
Myoid cells
B.
CELL PREPARATION
Mouse testis
Wildtype
Mutant
Dissociation
Mechanical
Enzymatic
Mechanical
Single-cell suspension
Total cell
FACS
Total cell
SEQUENCING: drop-seq
ANALYSIS: Pseudotime analysis, Tensor decomposition analysis, Cell clustering analysis
VALIDATION: Immunofluorescence
C.
Genes
Transcripts
Number of genes
Number of transcripts
SPG
SPCI
SPCII
SPD
WT–enzymatic
WT–mechanical
Mlh3–/–
Hormad1–/–
Cul4a–/–
CNP
D.
tSNE_1
tSNE_2
Undifferentiated spermatogonia
Differentiated spermatogonia
Leptotene/Zygotene
Pachytene
Meiotic division
Acrosomal
Round spermatids
Elongating spermatids
Sertoli
Leydig
Telocytes
Macrophages
E.
Wild-type
Mlh3-/-
Cul4a-/-
Hormad1-/-
CNP
F.
Wild-type
Mlh3-/-
Cul4a-/-
Hormad1-/-
CNP
G.
Wild-type
Mlh3-/-
Cul4a-/-
Hormad1-/-
CNP
Cell Type
Macro
Telo
Ley
Sert
eSpd
rSpd
Acr
MD
Pach
L/Z
dSpg
uSpg
Counts

Figure 1: Mapping cellular diversity in the adult testis using single-cell expression profiling. (**A**) Anatomy of the testis. Adult testes are comprised of germ cells (spermatogonia, primary spermatocytes, secondary spermatocytes, spermatids and spermatozoa) and somatic cells. Within the seminiferous tubules, there is a population of somatic cells (Sertoli). The tubules are wrapped by muscle-like "myoid" cells. Outside the tubules are a heterogeneous, poorly defined population of "interstitial" somatic cells including Leydig cells and telocytes. (**B**) Overview of the experiments. To establish the utility of single-cell profiling for testis phenotyping, we performed a series of experiments (i) comparing the quality of traditional enzymatic dissociation and more rapid mechanical dissociation, (ii) comparing the expression profiles of cells from total testis dissociation to testicular cells of known identity purified by FACS, (iii) comparing expression profiles of wildtype animals to cells isolated from 4 mutant strains with testis phenotypes (**Figure 1 - figure supplement 1**). (**C**) We used Drop-seq to profile 26,200 cells from wildtype animals and 31,400 cells from mutant animals, with an average of 1,155 transcripts/cell and 725 genes/cell (wildtype) and 2,223 transcripts/cell and 1,133 genes/cell (mutants). (**D**) We applied SDA and used t-SNE to visualize cells colored by k-means clustering of 20,322 cells, derived from our full dataset of wildtype and mutant animals, into 32 clusters (**Methods**, **Figure 1 - figure supplements 1- 5**). Label assignment clearly indicates a spatial organization of testis cells in t-SNE space, with somatic cell populations flanking the germ cells in small pockets. The full set of 32 clusters has been simplified into 12 major classes for ease of interpretation; the full clustering is shown in **Figure 1 - figure supplement 2**. (**E**) Histology sections from wildtype and mutant testis, illustrating the phenotypes observed in wildtype and the 4 mutant strains characterized by Drop-seq. Three of the strains, Mlh3-/-, Hormad1-/-and Cul4-/-have known pathology, while strain CNP represents an unpublished transgenic line with spontaneous male infertility. (**F**) Mapping of cells from each mouse strain into t-SNE space (colored points) compared to the background of all other strains (grey points). Mutant strains occupy distinct locations within t-SNE space, reflecting the absence of certain cell types in some strains (e.g. Mlh3-/-and Hormad1-/-), and alteration of expression in remaining cells (e.g. Hormad1-/-). (**G**)Counting individual cell types provides a quantitative phenotype of cellular heterogeneity in each strain.

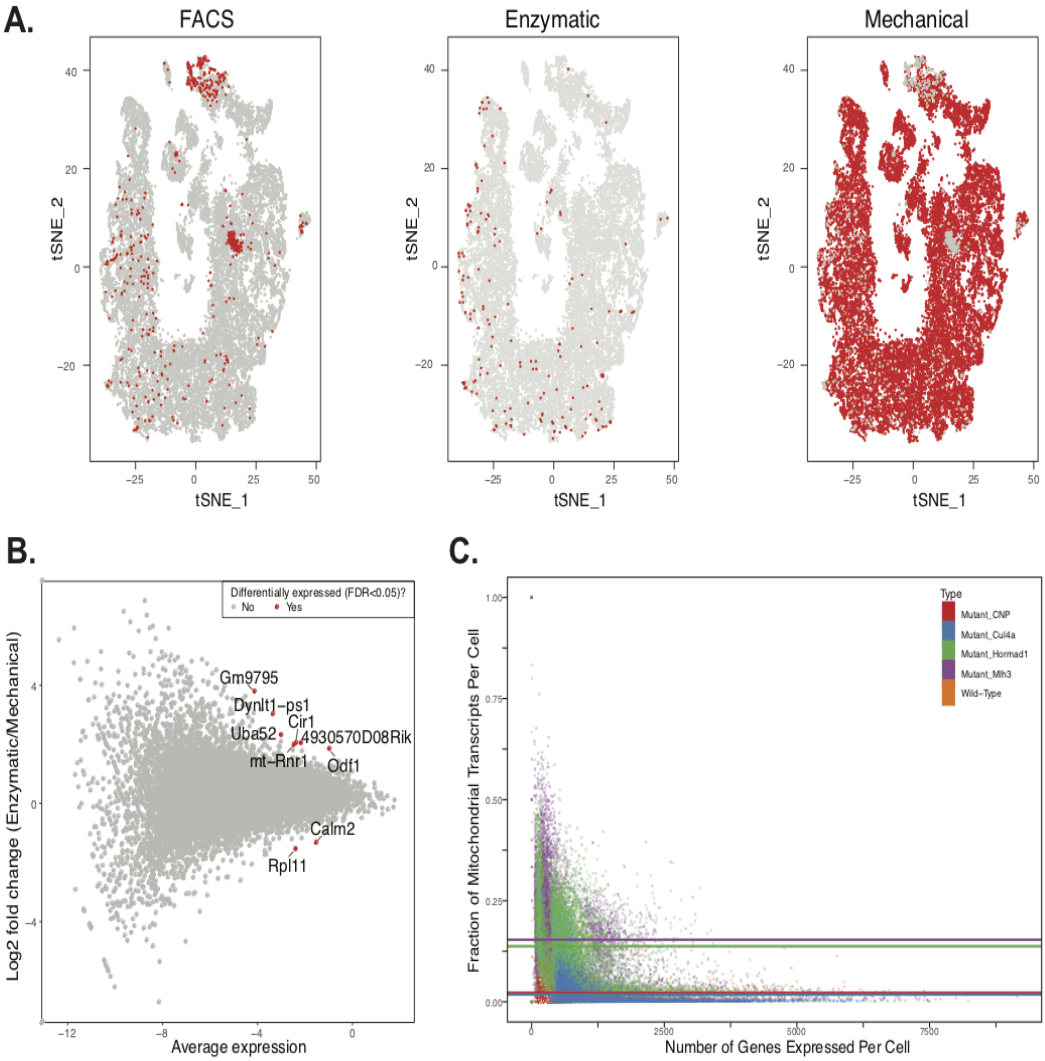


Figure 1 - Figure Supplement 1: Comparison of effects of dissociation protocols and mutation status on cell ascertainment and single-cell gene expression.
We compared the effects of enzymatic dissociation (SPG, WT1, WT2) and mechanical dissociation protocols (all other batches) on both the ascertainment of cell types (by visualization in tSNE space) and on single gene expression levels. (**A**) No obvious batch effects were detected when comparing the t-SNE clustering location of cells isolated by FACS, or either of the two total testis dissociation protocols. (**B**) We performed differential expression analysis to compare the two dissociation protocols, using all available WT datasets. We compared single gene expression levels for all genes detected by both protocols, summarized here as an M+A plot. The expression profiles from enzymatic and mechanical dissociation showed excellent concordance (R=0.95). Genes that were detected as differentially expressed by edgeR are plotted in red. (**C**) When compared to wildtype, cells from mutant strains exhibit significantly higher fractions of transcripts derived from mitochondrial genes, indicative of elevated rates of apoptosis.

Experiment	Dissociation Method	Sequencing Platform	STAMPS	Age	Strain	Bead Lot	Mapped Reads
SPG	Enzymatic	Hiseq	300	7 days	C57BL/6J	91615	41,388,604
SPCI	Mechanical	Miseq	200	2 months 21 days	C57BL/6J	91615	2,118,296
SPCII	Mechanical	Miseq	400	2 months 21 days	C57BL/6J	91615	2,055,974
SPD	Mechanical	Miseq	300	2 months 21 days	C57BL/6J	91615	1,978,523
WT1	Enzymatic	Miseq	200	3 months 21 days	C57BL/6J	91615	5,065,602
WT2	Enzymatic	Hiseq	600	4 months 2 days	C57BL/6J	91615	75,634,489
WT3	Mechanical	Miseq	1500	4 months	C57BL/6J	91615	8,816,148
WT4	Mechanical	Hiseq	3700	5 months 5 days	C57BL/6J	91615	44,120,247
WT5	Mechanical	Hiseq	2700	5 months 7 days	C57BL/6J	91615	50,515,490
WT6	Mechanical	Hiseq	3100	3 months 27 days	C57BL/6J	91615	53,237,395
WT7	Mechanical	Hiseq	2800	4 months	C57BL/6J	91615	61,540,944
WT8	Mechanical	Hiseq	2400	3 months 1 day	C57BL/6J	91615	65,841,512
WT9	Mechanical	Hiseq	3400	3 months 2 days	C57BL/6J	91615	58,792,345
WT10	Mechanical	Hiseq	2500	9 months 14 days	C57BL/6J	91615	45,875,980
WT11	Mechanical	Hiseq	2100	7 months 21 days	C57BL/6J	91615	65,055,410
Mlh3_1	Mechanical	Miseq	400	3 months 22 days	B6;129Mlh3-/-	91615	9,788,969
Mlh3_2	Mechanical	Miseq	300	8 months 30 days	B6;129Mlh3-/-	91615	19,791,310
Mlh3_3	Mechanical	Hlseq	800	8 months 30 days	B6;129 Mlh3-/-	91615	52,453,807
Mlh3_4	Mechanical	Hiseq	1500	2 months 21 days	B6;129 Mlh3-/-	91615	30,089,366
Mlh3_5	Mechanical	Hiseq	4300	2 months 22 days	B6;129 Mlh3-/-	91615	23,456,887
Mlh3_6	Mechanical	Hiseq	3900	2 months 23 days	B6;129 Mlh3-/-	91615	38,355,599
Hormad1_1	Mechanical	Hiseq	2300	4 months	B6;129 Hormad1-/-	72817	43,632,006
Hormad1_2	Mechanical	Hiseq	7700	4 months 3 days	B6;129 Hormad1-/-	72817	143,860,532
CNP_1	Mechanical	Hiseq	3400	1 month 18 days	C57BL/6J CNP-EGFP BAC TRAP	91615	17,649,845
CNP_2	Mechanical	Hiseq	1000	1 month 21 days	C57BL/6J CNP-EGFP BAC TRAP	91615	13,429,718
Cul4a_1	Mechanical	Hiseq	2800	2 months 14 days	B6;129 Cul4a-/-	91615	181,851,615
Cul4a_2	Mechanical	Hiseq	3000	2 months 16 days	B6;129 Cul4a-/-	91615	161,470,679
Total			**57600**				

Table 1. Summary of all wildtype and mutant single-cell RNA-sequencing experiments.

We added to this an additional 31,400 single cell profiles from 4 different mutant mouse strains exhibiting spermatogenesis defects: three mutants with known molecular mechanisms (knockouts of *Mlh3*, *Hormad1*, and *Cul4a*) as well as one knockin of a transgene (*Cnp*) that led to idiopathic infertility. We performed histological confirmation of testis defects in each animal prior to sequencing (**Figure 1E**). Seminiferous tubules in *Mlh3-/-* and *Hormad1-/-*mice exhibited a complete early meiotic arrest and absence of spermatozoa. *Cul4a-/-*sections showed a partial impairment of spermatogenesis, with a significant decrease in number of post-meiotic cells and abnormal spermatids. Sections from both *Cul4a-/-*and *Cnp* mice presented giant multinucleated cells but this type of cell was more prevalent in *Cnp* seminiferous tubules. Histological sections of *Cnp* mice displayed a clear defect in spermatogenesis as abnormal spermatids were observed;

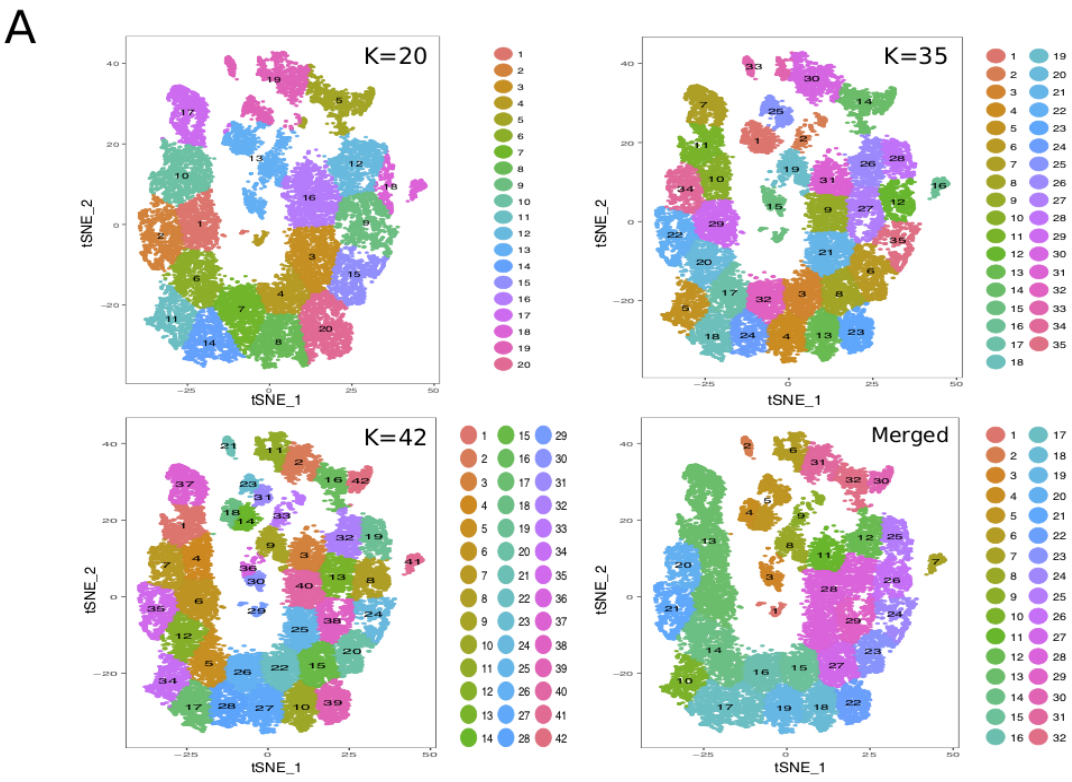
however, further molecular analysis is required to firmly characterize which stage(s) of spermatogenesis are affected.

3.3.2. Application of SDA, and comparison to classical clustering analysis

One specific challenge of analyzing a developmental system is that cluster-based cell type classification might artificially define hard thresholds in a more continuous process. Furthermore, a single cell's transcriptome is a mixture of multiple transcriptional programs, some of which may be shared across cell types. In order to identify these underlying transcriptional programs themselves rather than discrete cell types we applied SDA [88]. This is a model-based factor analysis method to decompose a (cell by gene expression) matrix into sparse, latent factors, or "components" that identify co-varying sets of genes which, for example, could arise due to transcription factor binding or batch effects (**Methods**). Each component is composed of two vectors of scores: one reflecting which genes are active in that component, and the other reflecting the relative activity of the component in each cell, which can vary continuously across cells, negating the need for clustering. This framework provides a unified approach to simultaneously soft cluster cells, identify co-expressed marker genes, and impute noisy gene expression (**Methods**). We inferred 50 components using SDA. Using these components, we visualized the overall results using t-distributed Stochastic Neighborhood Embedding (t-SNE) (**Methods**, **Figure 1D**): this t-SNE projection is also used in many subsequent analyses. We estimated the developmental ordering of cells using pseudotime modeling (**Methods**), based on our t-SNE embedding.

First, to provide a cross-check for our SDA results, we performed k-means (hard) clustering of our single cell libraries into discrete groups. (**Methods, Table 3**). We visualized the resulting

32 distinct clusters in tSNE space (**Methods**, **Figure 1D, Figure 1 - figure supplement 2**). Next, by inspecting the expression levels of known cell type markers and comparing to the FACS-sorted cells, we could resolve our 32 clusters into 11 distinct subtypes of germ cells and 4 somatic cell populations – Leydig cells, Sertoli cells, immune cells, and telocytes (**Figure 1 - figure supplement 2 and Figure 1 - figure supplement 3**). By tallying counts of cells within each cluster, we generated a digital readout of the cellular composition of wildtype and mutant animals (**Figure 1G, Figure 1 - figure supplement 4**), and are able to associate each SDA component to expression activity in particular cell type(s).


A
K=20
K=35
K=42
Merged
tSNE_1
tSNE_2

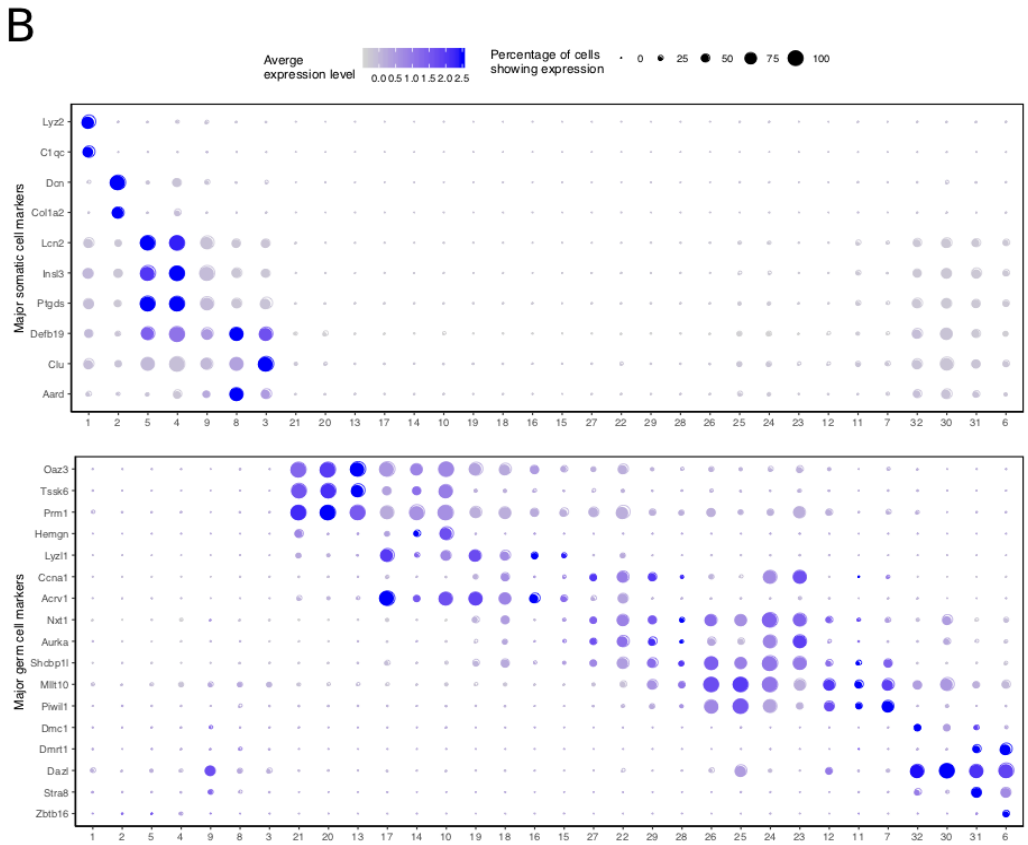

B
Averge expression level
0.0 0.5 1.0 1.5 2.0 2.5
Percentage of cells showing expression
0 25 50 75 100
Major somatic cell markers
Lyz2
C1qc
Dcn
Col1a2
Lcn2
Insl3
Ptgds
Defb19
Clu
Aard
Major germ cell markers
Oaz3
Tssk6
Prm1
Hemgn
Lyzl1
Ccna1
Acrv1
Nxt1
Aurka
Shcbp1l
Mlit10
Piwil1
Dmc1
Dmrt1
Dazl
Stra8
Zbtb16

Figure 1 - Figure Supplement 2: Mapping the Cellular Diversity of the Testis.
(**A**) We performed k-means clustering analysis of total joint wildtype and mutant cells with several levels of "k" to determine the appropriate level of clustering for cell type identification. Clusters with similar expression profiles were merged using out-of-bag-error (OOBE) method implemented in Seurat, ultimately leading to a final analysis with 32 clusters. (**B**) Expression of known somatic and germ cell--type specific markers in total joint wildtype and mutant cell clusters.

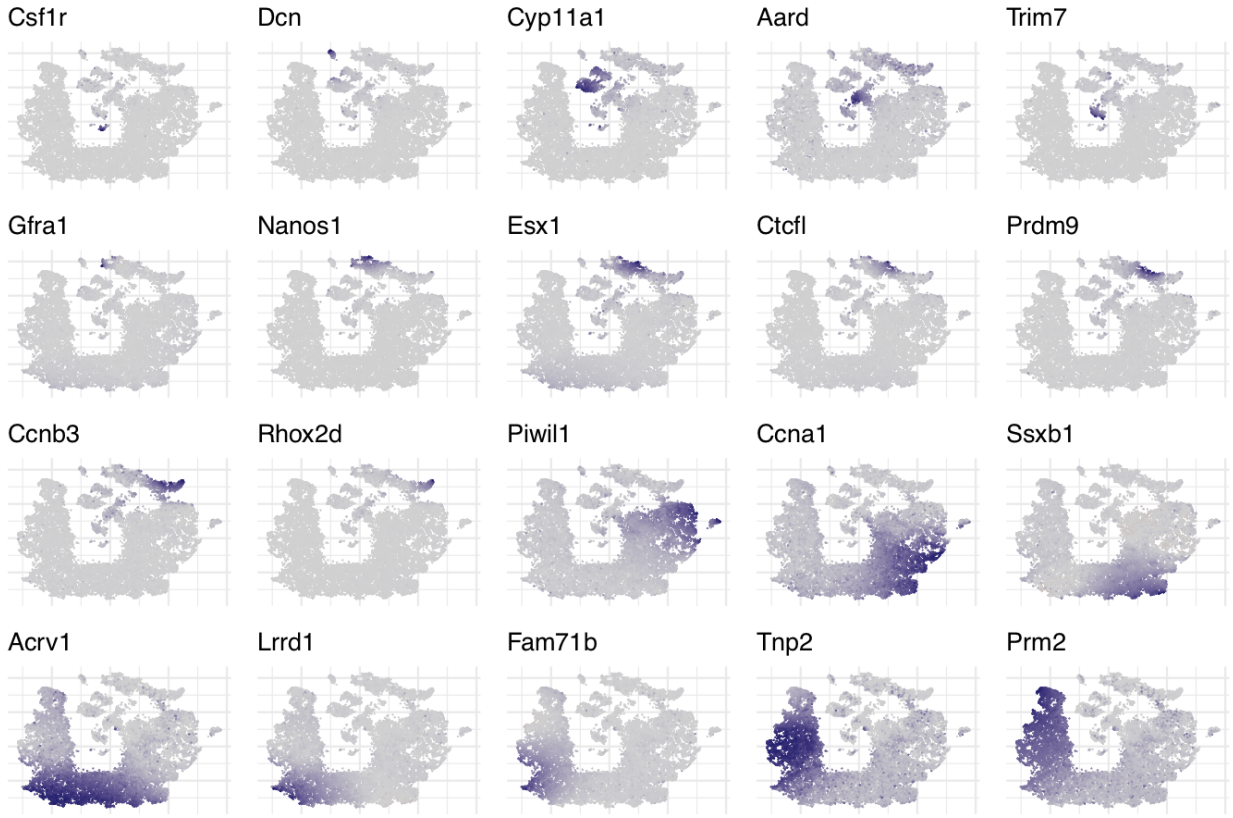
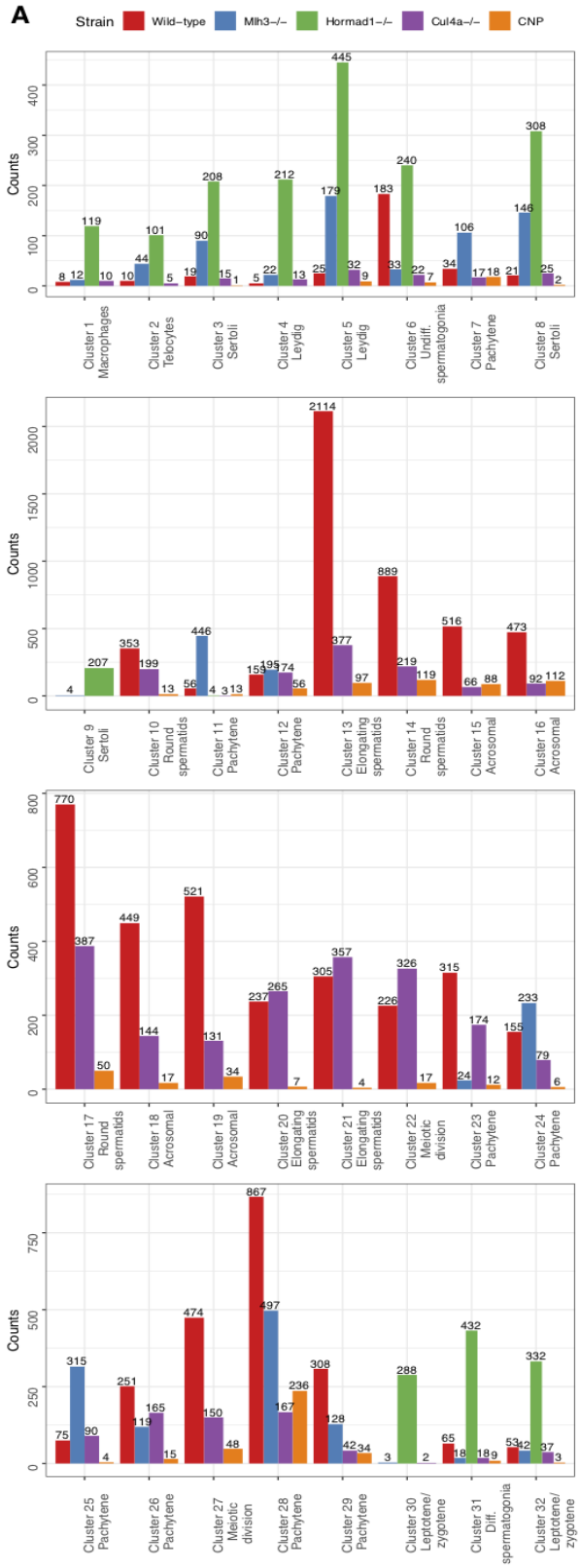


Figure 1 - Figure Supplement 3: Overview of expression patterns for some well-known testis cell markers in t-SNE space.

A



B

	Down−regulated	Not Significant	Up−regulated
Cluster 1	0	6584	0
Cluster 2	0	5253	0
Cluster 3	0	10114	1
Cluster 4	210	10461	23
Cluster 5	2	10408	9
Cluster 6	223	12127	166
Cluster 7	2	6916	0
Cluster 8	6	9817	13
Cluster 10	161	11259	35
Cluster 11	5	6733	3
Cluster 12	3	8966	0
Cluster 13	104	10906	4
Cluster 14	25	9082	13
Cluster 15	1	7442	1
Cluster 16	1	7121	1
Cluster 17	266	13259	22
Cluster 18	167	11123	22
Cluster 19	72	11218	7
Cluster 20	53	7686	17
Cluster 21	81	10029	21
Cluster 22	223	11965	21
Cluster 23	58	10370	33
Cluster 24	1260	11888	1
Cluster 25	162	12223	24
Cluster 26	120	11073	0
Cluster 27	6	8904	200
Cluster 28	44	10260	8
Cluster 29	10	8170	4
Cluster 31	103	10988	16
Cluster 32	697	9560	281

**Differential gene expression analysis beween wild-type and mutants for clusters 9 & 30 not present here as no wild-type cells are found in these clusters.*

Figure 1 - Figure Supplement 4: Tabulation of cluster counts by mouse strain and differential expression analysis within clusters.
(**A**) Count distribution of cells from each mouse strain for 32 clusters in **Figure 1 - Figure Supplement 2A** ("Merged"). (**B**) To explore any clustering bias between wild-type and mutants cells, we performed differential gene expression analysis using edgeR for each cluster between wild-type and mutant cells. Genes that did not express at least five transcripts were filtered. The cut-offs for significant differentially expressed genes were log-fold change of 1 and false discovery rate (FDR) adjusted p-value < 0.05.

Stage (Component[s])	**Seurat Cluster[s]**	**Example Key Genes**
Undifferentiated Spermatogonia (50 & 31)	6	Negative C50 loadings: Gfra1, Ccnd2, Glis3, Zfp462, Tex19.1, Dppa4 C50 loadings close to 0: Zbtb16 aka Plzf, Sox4, Afp, Mageb4, Foxo1 Positive C50 loadings: Nanos3, Lin28a, Foxf1, Pramef12, Sox3 (All have positive C31 loadings)
Differentiating Spermatogonia (7)	6	Glis2, Nanos1, Rcor2, Zswim5
Spermatogonia (Broad) (33)	6	Uchl1, Dmrt1, Sohlh1, Dnmt3a, Dnmt3b, Dnmt1, Scml2, Msh2, Map7d2, Ung
Intermediate/B Spermatogonia (2)	31	Ctcfl, Esx1, Pou4f1 aka Brn-3a, Tex13b

(Pre)Leptotene (5)	32	DSB associated: Prdm9, Setdb1, Dmc1, Gm960 (aka Top6bl), Brca2, Tex15, Ddb2, Brip1, Msh5, Mms22l, Meilb2 (Hsf2bp), Mcm8, Rad51, Ccdc36 (Iho1) ZMM: Shoc1 (AI481877), Msh5, Brip1 Cohesin & synaptonemal components: Rad21l, Smc1b, Smc3, 4930447C04Rik (aka Six6os1), Tex12 Ctcfl regulated: Prss50, Stra8, Ugt8a, Gal3st1 Telomere tethering: Terb1, Terb2
Zygotene (44)	32	Rad51ap2, Meiob, Spata22, Hfm1
Early Pachytene (13)	24, 25, 26	Meiotic cell cycle: Ccna1, Ccnb3, Aurka, Plk1 piRNA associated [Better known drosophila homologues in square brackets]: Piwil1 [Miwi], Tdrd1 [Tudor], Tdrd5 [Tejas], Pld6 [Zucchini] Protein folding: Hspa5, Calr, Hsp90b Fertilization: Zpbp, Zpbp2,
Early Pachytene 2 (47)	24, 25, 26	Chromosome function: Hormad1, Setx, Ncaph, Kdm1b Spindle function: Cenpe, Cntrob, Pcm1 Meiotic cell cycle: Stambp, Ccnb1ip1, Ccnb3, Gfra4

Mid Pachytene (48)	7, 24, 25	Cilium/axoneme assembly: Cfap46, Cfap65, Cfap74, Dnah2, Dnah12, Dnah14, Dnhd1, Ak7, Ccdc39, Mroh2a Microtubule/spindle function: Dcdc2b, Ccdc88a, Knl1 Splicing: Srrm2, Tra2a, Srek1, Rbm5, Rbm25
Pachytene and late pachytene (42, 39)	22, 24, 26	piRNA associated [Better known drosophila homologues in square brackets]: Piwil1 [Miwi], Tdrd1 [Tudor], Tdrd5 [Tejas], Tdrd12 [Yb], Piwil2 [Mili], Mael [Maelstrom], Pld6 [Zucchini], Exd1 [Egalitarian], Ddx4 aka Mvh [Vasa], Tdrkh [Papi], Tdrd9 [SpnE] Meiotic cell cycle: Calm1, Calm2, Calm3, Meig1, Lyar, Spata4, Cetn1, Mns1 Translational Repression: Ybx1 (aka MSY2), Ybx3, Pabpc1 Cilium assembly: Rsph1, Ropn1l, Dnah8, Dnaaf1, Cfap36, Bbof1, Ccdc39 Post-meiotic (fertilization and metabolism): Ldhc, Dkkl, Clgn, Spink2, Catsperz, Fbp, Cct1, Cct3, Cct4, Cct7
Diplotene & Divisions (20)	22	Cell cycle: Ccnb1, Ccna1, Cdc25a, Aurka, Bora, Plk1, Rgcc, Fzr1 DUF622 containing: 1700001F09Rik, Gm3453, Gm10354, Gm3149, Gm8362, Gm3127, Gm17019, Gm4181, Speer4e, Speer4b, Gm9758, Gm8232, BC061237, Gm5458, 4930572O03Rik, Gm5800, Gm7361, Gm8220 SSXRD and KRAB-related domain containing: Ssxb1, Ssxb2, Ssxb3, Ssxb5, Ssxb6 Others: Tbpl1 (aka Tlf)

Spermatid - Acrosome (30)	17, 19	Acrosomal: Spaca1, Spaca3 (aka Lyzl3), Spaca4, Spaca5 (aka Lyzl5), Spaca7; Lyzl1, Lyzl4, Lyzl4os, Lyzl6; Acrv1, Aep1, Spata9, Spata31, & Spata46 Sperm-oocyte interaction: Izumo1, Izumo3, Zpbp, Zp3r Others: Catsper1, Catsper3, & Catsper4; Tekt1, Tekt2, Tekt3, & Tekt4; Creb3l4 aka Atce1, and 1700016D06Rik + Lrcc34
Spermatid - Mysterious (35)	17, 10	Testis enriched genes of unknown function: Tex29, Lrrd1, Smco4, Heatr9, Hsfy2, Tepp, Spata31d1d, Tmem81, Spata25 Mitochondrial function: Crls1, Slc25a41
Spermiogenesis (17 [& 18, 34])	13, 20, 21	Histone Replacements: Prm1, Prm2, Prm3, Tnp1, Tnp2 Others: Smcp, Odf2, Gapdhs, Oaz3, H1fnt (aka H1t2), Pgk2, and Cabs1 4+ Abhd5
Leydig (40)	4, 5	Testosterone Biosynthesis: Star, Cyp11a1, Hsd3b1, Cyp17a1, Hsd17b3 Others: Insl3, Ptgds
Sertoli (45)	8, 3	Aard, Defb36, Cst12, Ldhb, Tmsb4x, Cst9, Gstm6, Sin3b, Gsta4, Chchd10, Gstm7, Basp1, Wfdc10
Macrophages (11)	1	Csf1r, Cd163, Cd68, Adgre1 (aka F4/80), Itgam (aka CD11b), Mrc1, Cx3cr1, Fcgr3, C1qa, C1qb, C1qc
T-cells (3)	1	Ptprc (aka CD45), Il2rg, Cd3g, Cd3d, Cd3e, Trbc2, Trac, Ms4a4b, and Cd2
Telocyte (32)	2	Dcn, Cd34, Pdgfra, Col1a2, Col3a1, Col6a1, Col4a4, Col4a1, Col1a1, Lamb1, Lama2, Lamb2, Des

Peritubular Myoid (21)	3	Cnn3, Vegfa, Edn1
Hormad1 KO (38)	30	X & Y linked: Zfy1, Zfy2, Rhox2h, Rhox2d, Rhox2a, Rhox2c, Rhox2g Autosomal: Dnajc12, A830018L16Rik (aka C8orf34)
Cul4a KO (25)	17, 10	Hist1h2al, Csmd1, Jakmip2, Tagln2, Map2k7, Lpo
Respiration (9)	22, 23, 26, 27	Complex I (NADH:ubiquinone oxidoreductase) Ndufa11, Ndufa12, Ndufa2, Ndufa3, Ndufa5, Ndufa6, Ndufa7, Ndufaf2, Ndufaf8, Ndufb10, Ndufb2, Ndufb4, Ndufb5, Ndufb7, Ndufb9, Ndufc1, Ndufs4, Ndufs6, Ndufv3 Complex III (Ubiquinol-cytochrome c reductase) Uqcc2, Uqcr10, Uqcr11, Uqcrb, Uqcrh, Uqcrq Complex IV (Cytochrome c oxidase) subunits Cox17, Cox4i1, Cox5a, Cox5b, Cox6a1, Cox6b2, Cox6c, Cox7a2, Cox7b2, Cox7c, Cox8c Cytochrome c: Cyct Complex V (ATP Synthase) Atp5e, Atp5h, Atp5j, Atp5j2, Atp5k, Atpif1
Batch Effect (22)	17, 19, 28	(all downregulated) Ribosomal Proteins: Rps7, Rpl11, Rps13, Rps12, Rps17, Rps23, Rpl18a, Fau (Rps30 fusion) Others: Tpt1, Kpna2, Eif1
Batch Effect (12)	31, 6	Gm42418, Rbm25, mt-Rnr1, mt-Rnr2, Ncl, Pet2, Vps8

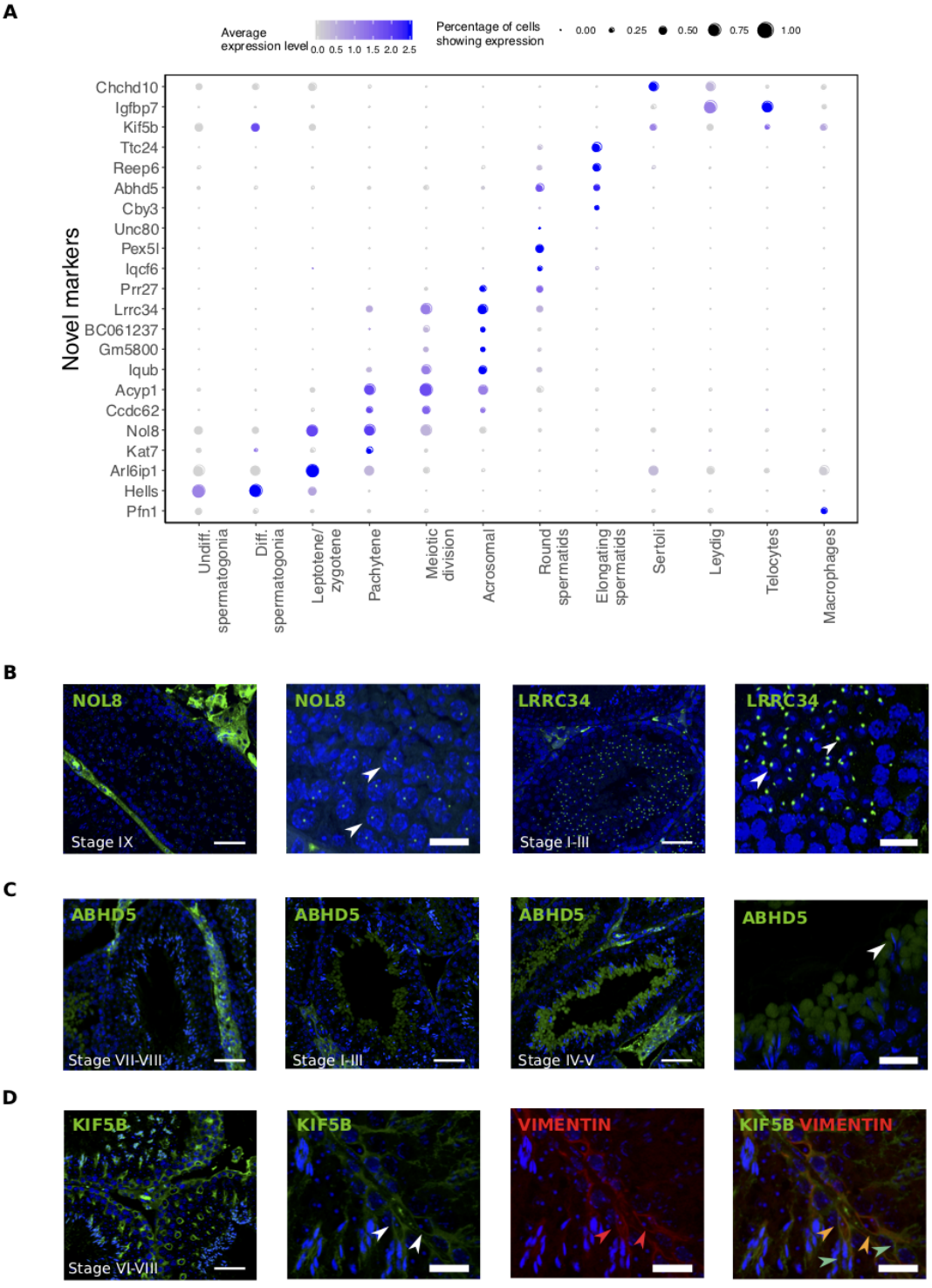
Table 3. Key genes from example SDA components of different stages

Careful examination and quantification of cell-type composition differences in each mutant strain recapitulated the known pathology of mutants (*Mlh3-/-*, *Hormad1-/-*and *Cul4a-/-*) at digital resolution. The location of mutant cells in t-SNE space illustrated the absence of certain cell types within spermatogenesis (**Figure 1F)**. Consistent with the known biology, we observed that both *Mlh3-/-*and *Hormad1-/-*cells arrest at different stages of meiosis I; mid-pachytene and leptotene/zygotene respectively. Derangement of certain cell types in the developmental trajectory was also observed as leptotene/zygotene *Hormad1-/-*cells formed distinct clusters. Both t-SNE and hard clusters indicated strong mixing of mutant and wildtype cells; of the 32 clusters, only 2 did not contain both wildtype and mutant cells. Both lacked wildtype cells: cluster 9, a Sertoli cell cluster, and cluster 30, containing leptotene spermatocytes primarily from *Hormad1-/-*. As the bulk of our experiments were performed on total testis samples, we do see preferential ascertainment of some cell types from the mutant strains depleted of post-meiotic germ cells: 95% of somatic cells (clusters 1-5,8,9) and 83% of pre-pachytene germ cells (clusters 6, 30-32) are derived from mutants (**Figure 1 - figure supplement 4A**). The majority of these clusters have fewer than 10 genes with differential expression detectable between mutant and wildtype (**Figure 1 - figure supplement 4B**), and we proceeded with a joint analysis of mutant and wildtype cells, with the caveat that conclusions about the biology of these particular clusters are derived largely from mutant strains.

3.3.3. New molecular markers of cellular subtypes

Single cell RNA sequencing provides new opportunities to assess important open questions in the field of spermatogenesis. Along with the expected patterns of expression for known markers, we identified numerous novel markers for all populations, some of which we selected for

validation using immunohistochemistry (**Figure 2**). Noteworthy are the identification of KIF5B as a Sertoli cell protein that provides more extensive coverage of the cell body than the conventional markers TUBB and VIM, and the identification of ABHD5 as a marker for the subcellular structure of developing germ cells known as the residual body. Protein products for predicted markers ACYP1, UNC80, and CCDC62 were not detected, which might be an antibody-related problem or an indication that these RNAs were not translated. (**Figure 2**).


A
Average expression level
0.0 0.5 1.0 1.5 2.0 2.5
Percentage of cells showing expression
0.00 0.25 0.50 0.75 1.00
Novel markers
Chchd10
Igfbp7
Kif5b
Ttc24
Reep6
Abhd5
Cby3
Unc80
Pex5l
Iqcf6
Prr27
Lrrc34
BC061237
Gm5800
Iqub
Acyp1
Ccdc62
Nol8
Kat7
Arl6ip1
Hells
Pfn1
Undiff. spermatogonia
Diff. spermatogonia
Leptotene/ zygotene
Pachytene
Meiotic division
Acrosomal
Round spermatids
Elongating spermatids
Sertoli
Leydig
Telocytes
Macrophages
B
NOL8
Stage IX
NOL8
LRRC34
Stage I-III
LRRC34
C
ABHD5
Stage VII-VIII
ABHD5
Stage I-III
ABHD5
Stage IV-V
ABHD5
D
KIF5B
Stage VI-VIII
KIF5B
VIMENTIN
KIF5B VIMENTIN

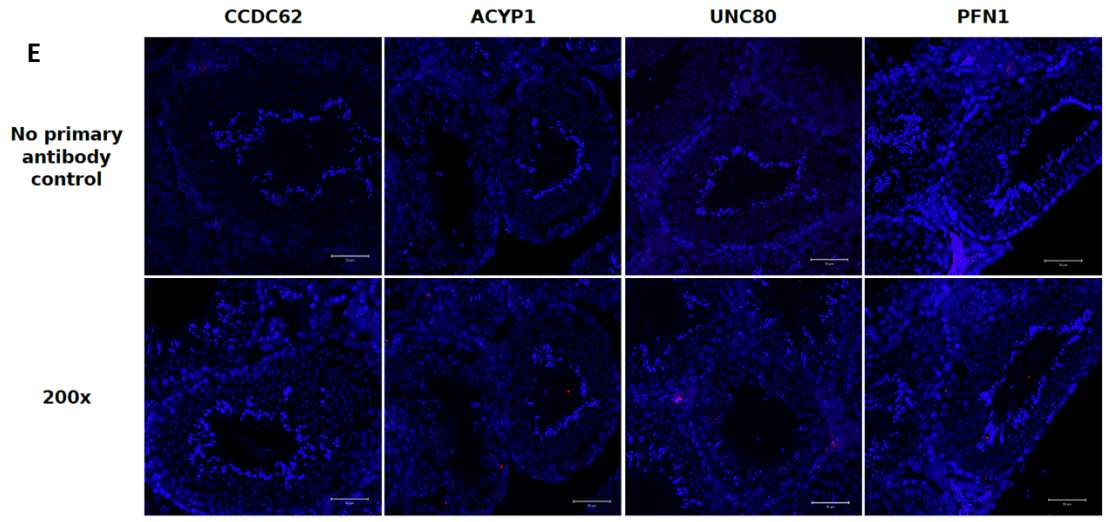


Figure 2: Identification of novel cellular markers from single-cell data.
(A) Across major cell-type clusters, we identified 22 gene expression markers specific to one cell type or aspect of spermatogenesis and not previously reported. Here we show the expression levels of these genes. Expected protein expression patterns for Nol8, Lrrc34, Abhd5, and Kif5b were confirmed, but the antibodies for Acyp1, Ccdc62, and Unc80 did not show positive staining in any testicular cell types, which could be an antibody-related problem or an indication that these RNAs were not translated. (B-E) Thin scale bar, 50 µm; thick scale bar, 20 µm. (B) Nol8, a nucleolar protein, marks primary spermatocytes while Lrrc34marks nucleoli in round spermatids (white arrowheads) (C) Within the tubules, Abhd5 marks specific cytoplasmic regions of elongating spermatids destined to form the residual body (white arrow head) and staining intensity peaks during seminiferous tubule stages IV-V. (D)Kif5b marks Sertoli cells within seminiferous tubules (white arrow head). We co-stained Kif5b with a well-known Sertoli cell marker, Vimentin (red arrow head), and indeed both proteins colocalize to Sertoli cells (orange arrow head). Co-staining also reveals that Kif5b staining extends further out in the cell body (blue-green arrow head) than Vimentin.

We identified a number of somatic cell populations (hard clusters 1,2,3,4,5, 8 and 9 in **Figure 1 Supplement 2A "Merged"**). Because our SDA analysis suggested multiple components, varying even within these clusters (**see below**), we performed additional targeted hard clustering analyses on these cells **(Methods**), identifying additional complexity: 10 identified somatic cell clusters comprise 4 Sertoli cell sub-clusters, 3 Leydig sub-clusters, 2 immune cell clusters (macrophages and lymphocytes) and 1 telocyte cluster (**Figure 1 - figure supplement 5**).Telocytes are a recently reported stromal cell type present in a wide range of

tissues, and are little studied in testis [89]. In addition to the previously reported markers Cd34 and Pdgfrα, we find a number of even more highly specific expression markers for telocytes, including Dcn, Gsn, Tcf21 (**Table 2**).

https://wustl.box.com/s/7klp1yp5qcmcgw1y24hj6zemzgov80g6

Table 2. Summary of all differentially expressed genes in total joint wildtype and mutant cell clusters.

All Sertoli cell sub-clusters express known Sertoli cell-type specific markers (**Figure 1 - figure supplement 5C**) and these Sertoli sub-clusters are enriched with different GO terms (biological processes) including cytoskeleton organization (sub-cluster 1), protein folding (sub-cluster 2), RNA splicing (sub-cluster 2 and 3) and spermatogenesis (sub-cluster 4) **Figure 1 - figure supplement 5D**). All Leydig sub-clusters express known Leydig-specific genes (**Figure 1 - figure supplement 5C**) and these Leydig sub-clusters are marked by different GO (biological processes) terms which include steroid and lipid biosynthetic process (sub-cluster 1), ATP synthesis coupled electron transport and drug metabolic process (sub-cluster 2) and cofactor and steroid metabolic process (sub-cluster 3) (**Figure 1 - figure supplement 5D**).

3.3.4. SDA-based gene expression modules

Based on the above analyses, it is clear that our 50 SDA components represent all the major different cell types and developmental stages of spermatogenesis, with other specific components capturing batch effects and general processes such as respiration. Encouragingly, most components contained relatively few highly expressed genes (**Figure 3B)** and, when compared to alternative commonly used methods for matrix factorization (non-negative matrix factorization, NNMF; principal component analysis, PCA; independent component analysis, ICA), SDA produced the most sparse model (**Figure 3C**).Although the model used for inference is symmetric for

positive/negative gene weightings, many identified components showed strong biases towards positive or negative weights, consistent with expectations for identifying a group of co-activated (or co-repressed) genes (e.g. **Figure 3E**). Likewise, the cell loadings of each component frequently highlight specific cellular subsets that localize in tSNE space and pseudotime (**Figures 3D&F, Figure 3 - figure supplement 1**), and often interpretable as particular identifiable cell types in our initial hard clustering. Thus, we label SDA components as "expression modules". We found that most components generated from an SDA analysis of only wildtype data were also observed in the joint analysis of wildtype and mutant data, which we proceed to use for the remaining analyses (**Figure 3 - figure supplements 2-3**).

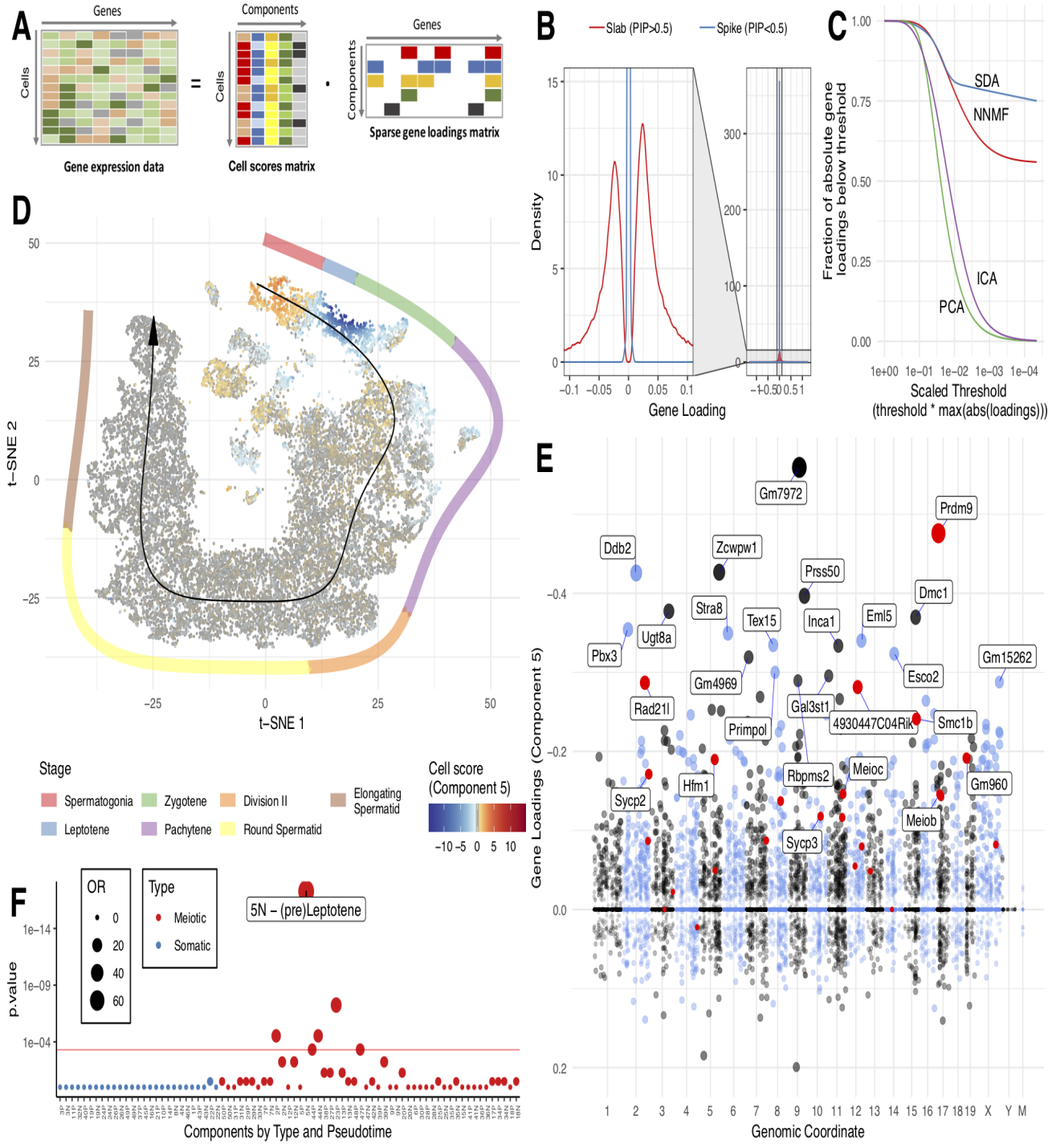


Figure 3: SDA identifies gene modules and maps them to cells.
(A)We applied sparse decomposition analysis (SDA) to identify latent factors ("components") representing gene modules. These components are defined by two vectors – one that indicates the loading of each cell on the component, and one that indicates the loading of each gene on the

component. (B) SDA uses a spike and slab prior on the gene loadings to induce sparsity (a point mass at 0 and a centered normal distribution respectively). PIP=Posterior Inclusion Probability that a gene loading is not equal to zero (i.e. not in the spike). The figure shows the density of gene loadings over all components with loadings separated into genes with PIPs >0.5 (20%) versus <0.5, indicating the sparsity of resulting gene loadings. (C) SDA produces sparser representation of gene loadings compared to other matrix factorizations: NNMF, ICA and PCA. For each method, the fraction of all absolute gene loadings exceeding a "no loading" sparsity threshold is shown, normalized by the maximum absolute loading across all components for that method. (D)We fitted 50 SDA components using 20,322 wildtype and KO cells (see also Figure 3 - figure supplements 1-5). We illustrate component 5. The loadings of component 5 in t-SNE space highlight a cluster of cells at the leptotene early meiotic developmental stage. Black arrow: the principle curve fit to the germ cell data, corresponding to the developmental ordering of each cell progressing through spermatogenesis. The colored segmented line shows broad staging of spermatogenesis. (E)Genomic location versus loadings for component 5. Most genes have near-zero loadings, but a fraction have non-zero loadings, including the well-known histone methyltransferase Prdm9. Red genes: GWAS hits for human recombination rate. (F)Component 5 is highly and specifically enriched for GWAS hits of human recombination rate. OR: Odds Ratio. P value by FET (main text). Positive (P) and negative (N) loadings are tested separately. For one sided components (cell score range ratio >5) the minor side is omitted. Red horizontal line: p=0.05 after Bonferroni correction for multiple testing.

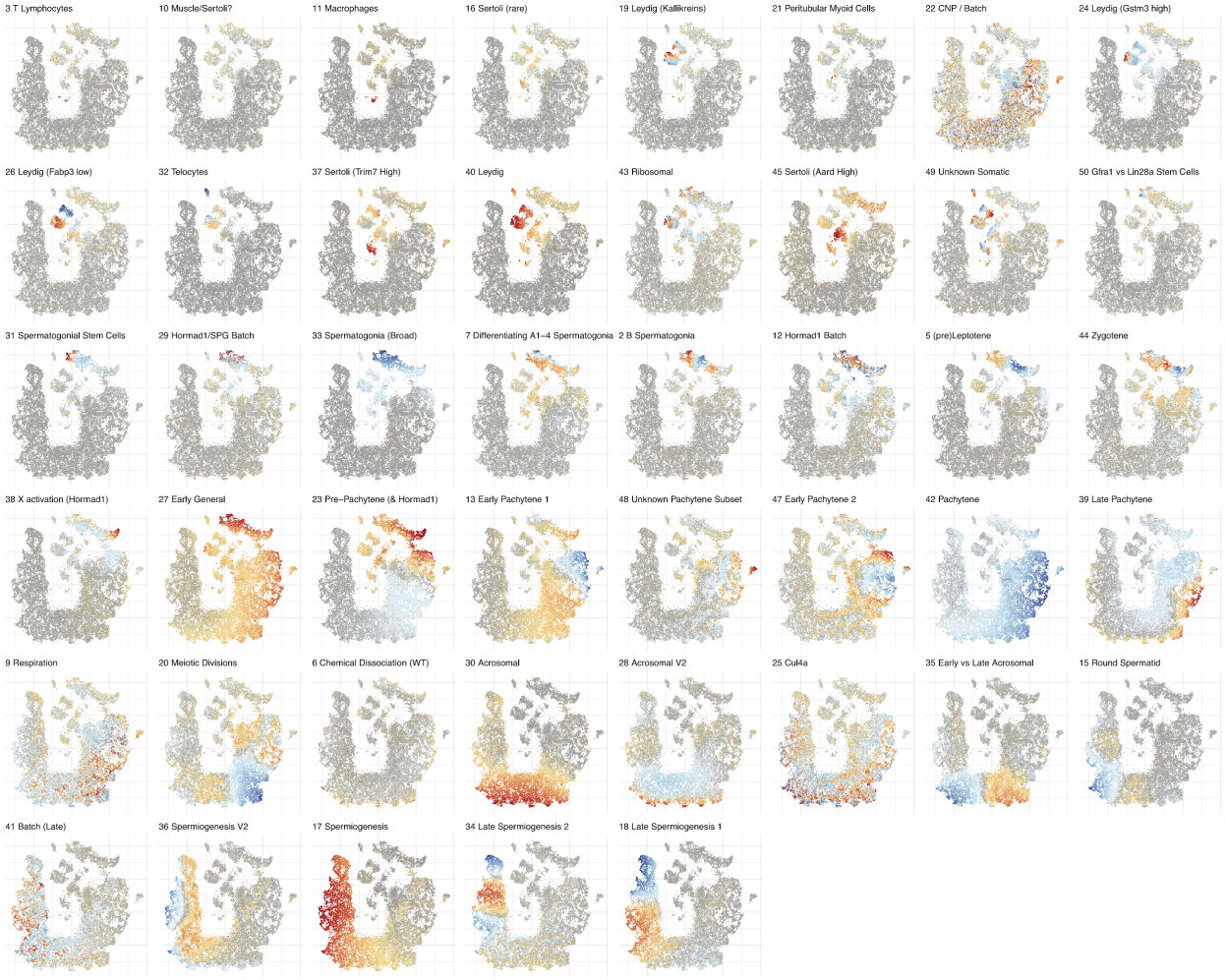


Figure 3 - Figure Supplement 1: Overview of cell score loadings in t-SNE space for all components produced by SDA except single cell components (1, 4, 8, 14, & 46).
The component number and biological interpretation of the component are labelled above each panel.

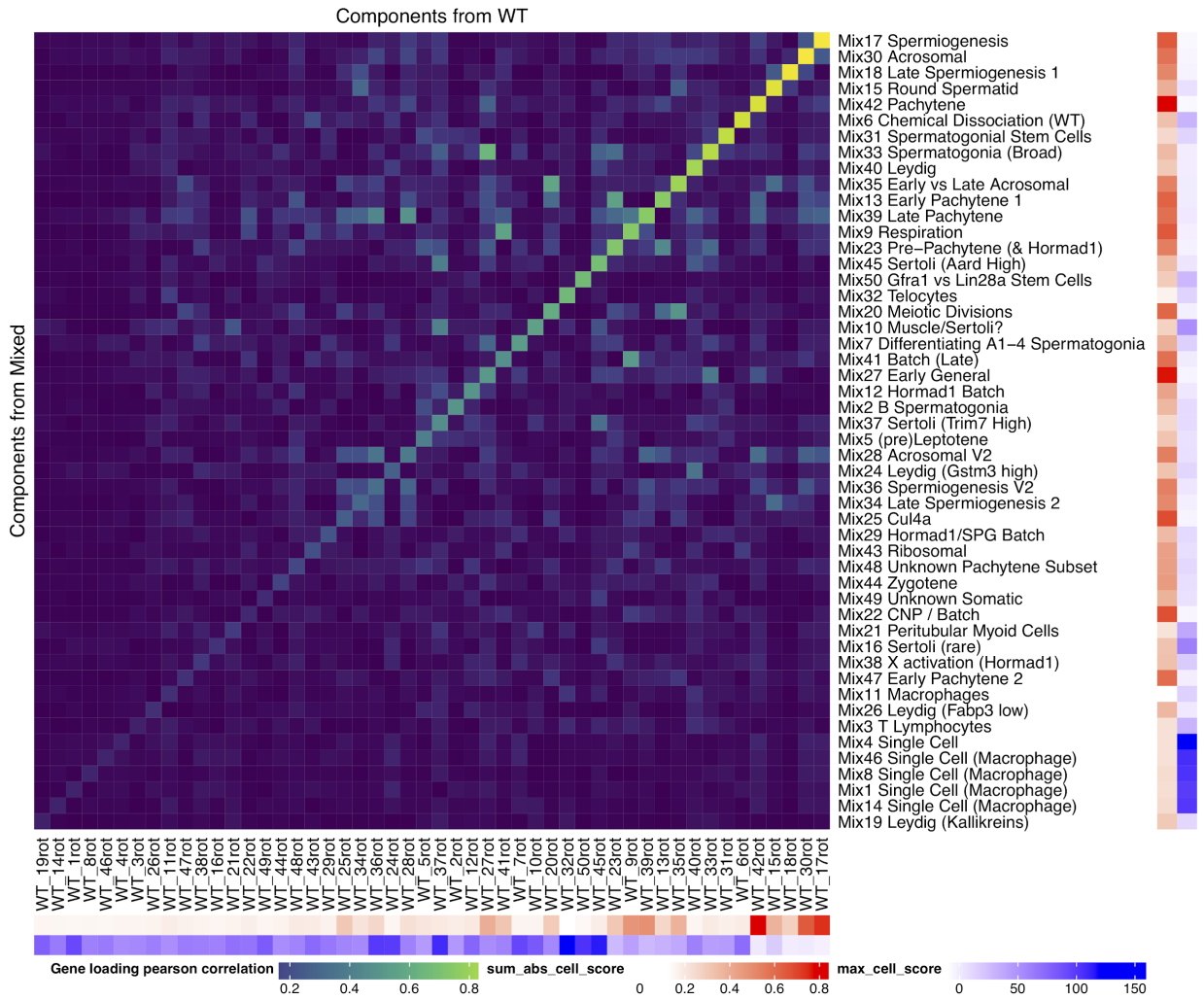


Figure 3 - Figure Supplement 2: Robustness of SDA Results.
All of the SDA results presented in the main text are derived from a combined analysis of all wildtype and mutant cells (the "Mixed" analysis). In order to quantify the robustness of our conclusions to this decision to combine mutant and wildtype strains, we performed a separate SDA analysis using just wildtype cells (the "WT" analysis). Here we show as a heatmap the pearson correlation of component gene loadings between a procrustean rotation of the WT gene loadings and the Mixed SDA gene loadings. The "sum abs cell score" annotation shows the sum of the absolute cell scores for that component (larger number indicates a more important component). The "max cell score" annotation indicates the maximum cell score for each component (a larger maximum indicates overfitting to a single / small number of cells). The most important WT components have high correlations with components in the mixed SDA run. Some components such as Mix38 X activation do not appear in the WT decomposition because they represent mutant specific processes. Other components such as Mix44 Leptotene-Zygotene do not appear as these cells are enriched in mutant samples due to the lack of later cells.

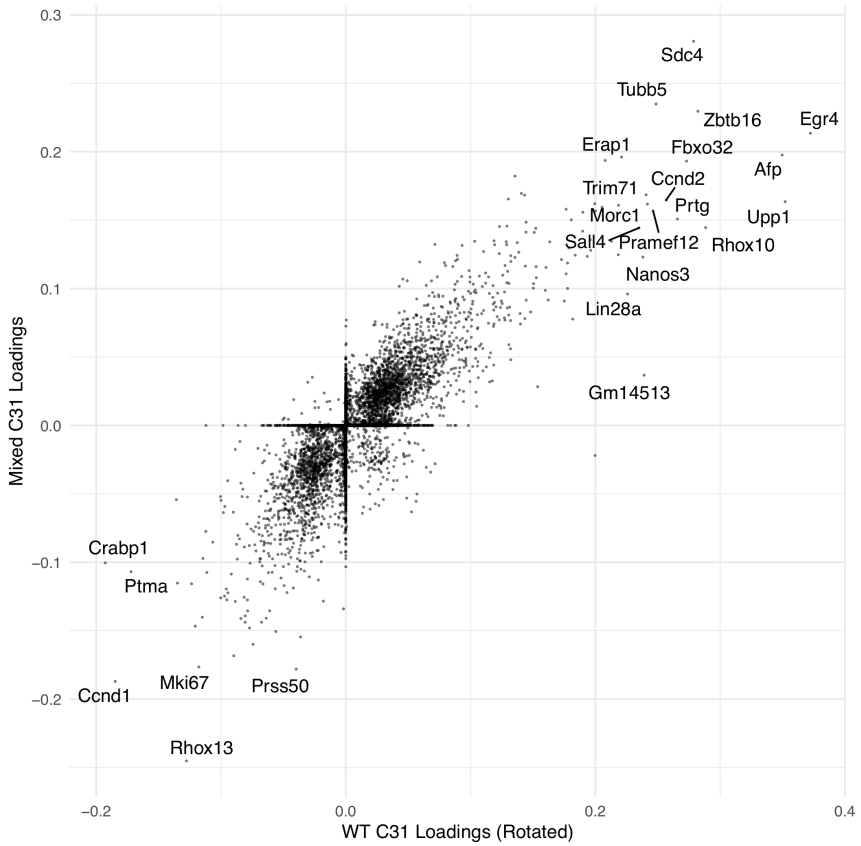


Figure 3 - Figure Supplement 3: Correlation of C31 gene loadings.
An example scatterplot comparing the gene loadings for one cognate SDA component (C31) between WT and Mixed SDA runs. The correlation is high.

To provide further intuition towards how SDA components summarize transcriptional programs, we selected 14 components that, collectively, load highly on germ cells throughout spermatogenesis. When we visualize the total expression output for each cell, ordered by pseudotime, as a sum of all 14 components, it is clear that expression can be modeled as an overlapping series of components in time, coming on and going off gradually over different timescales (**Figure 4A-B**). Each component is enriched for specific genes, and, importantly, genes with different identified biological functions (**Figure 4C&D**). SDA components provide complementary information to hard clustering: a single hard cluster may have significant cell scores from as many as three components, indicating multiple different expression programs jointly active in each cell. Conversely a single SDA component may show significant cell scores

across more than three hard clusters (**Table 3**), emphasizing that expression changes gradually as cell types and fates evolve (**Figure 4 - figure supplements 1-2**).

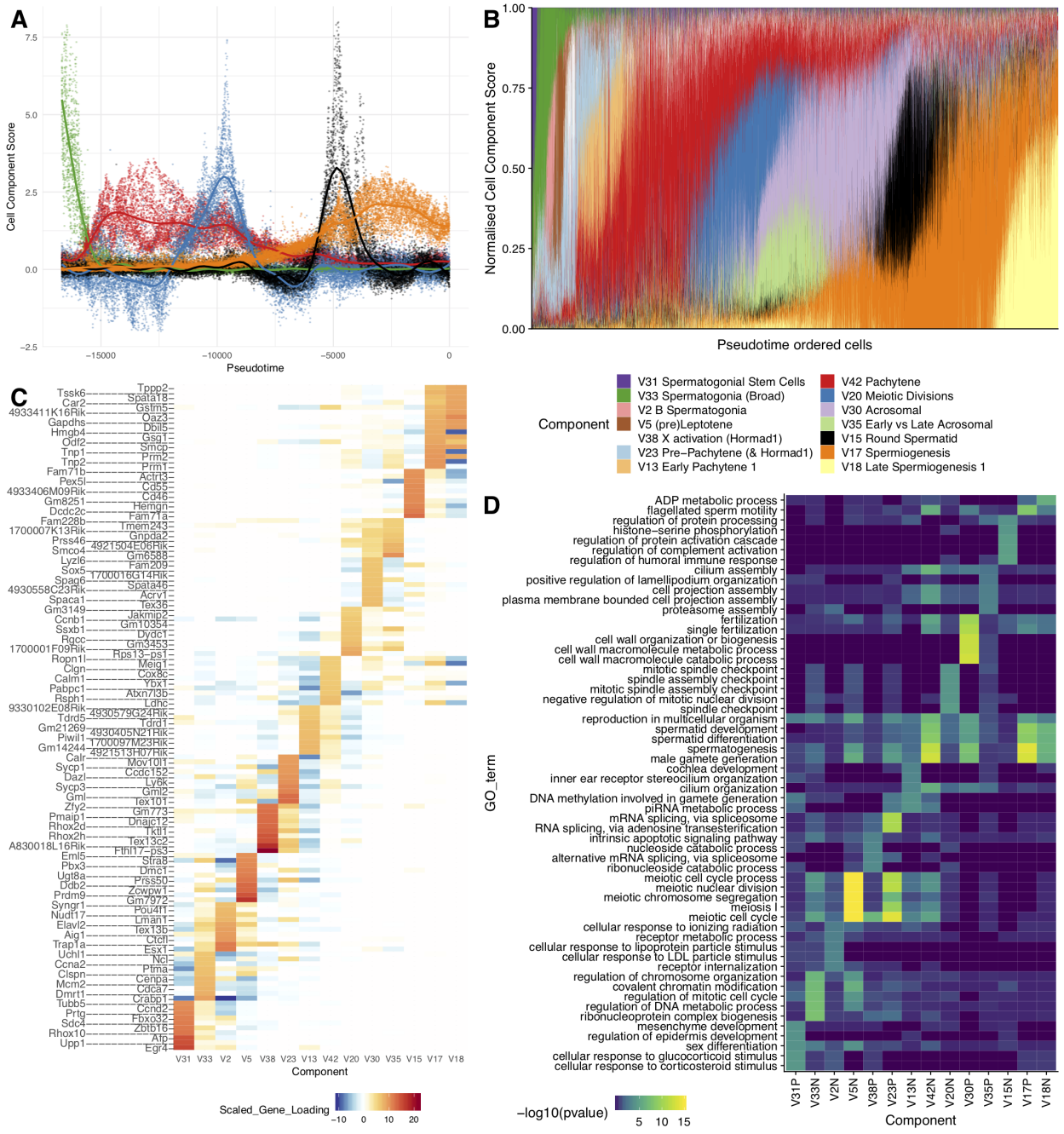


Figure 4: SDA components overlap but represent distinct processes.
(A) For five example components, the cell scores for each cell are plotted through pseudotime, indicating strongly overlapping dynamically varying component activity. Component signs were

chosen to be mainly positive (components have arbitrary sign). Color mappings as in panel B. **(B)**Stacked bar plot of cell component loadings for 14 germ components sorted by cell pseudotime. Each column corresponds to an individual cell and the total positive component loadings for each are normalized to 1 after flipping components to be mainly positive. Factorization by SDA indicates that transcription during spermatogenesis can be represented as an overlapping series of components in time, coming on and off gradually on different timescales. See also **Figure 4 - figure supplements 1-2** for alternative visualizations of components in pseudotime. **(C)**Furthermore these components are comprised of distinct gene sets driving distinct biological processes. Shown are the top 10 gene loadings for each of the components in**(B)** represented as a heatmap. Most genes have strong loading on only one component. **(D)**Likewise, a gene ontology enrichment analysis for biological processes in the top 250 genes for each component indicates largely non-overlapping enrichments across components. More in-depth analysis of GO enrichments and gene loadings for each component allow separation of components into biological and technical effects (**Figure 4 - figure supplements 3-4**).

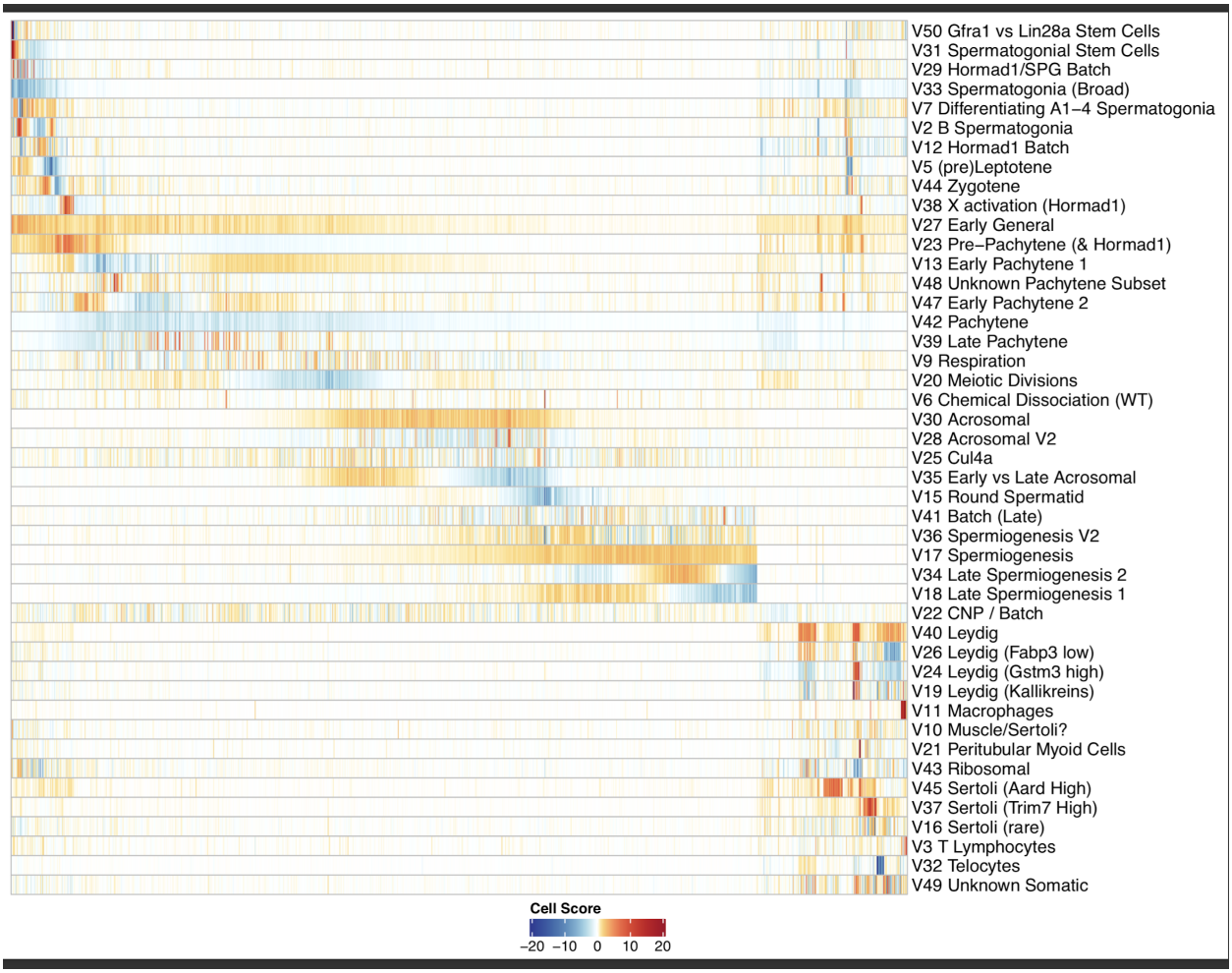


Figure 4 - Figure Supplement 1: Heatmap of SDA component scores.

Cell scores for all SDA components except single cell components. Each column corresponds to an individual cell and each row is a component. The columns and rows are both ordered by pseudotime, except for the somatic cells/components in which the components are ordered alphabetically and cells are ordered by cluster label assigned by hierarchical clustering applied to all cells (method ward.D2, k=25). Absolute cell scores greater than 20 have been rounded down for plotting.

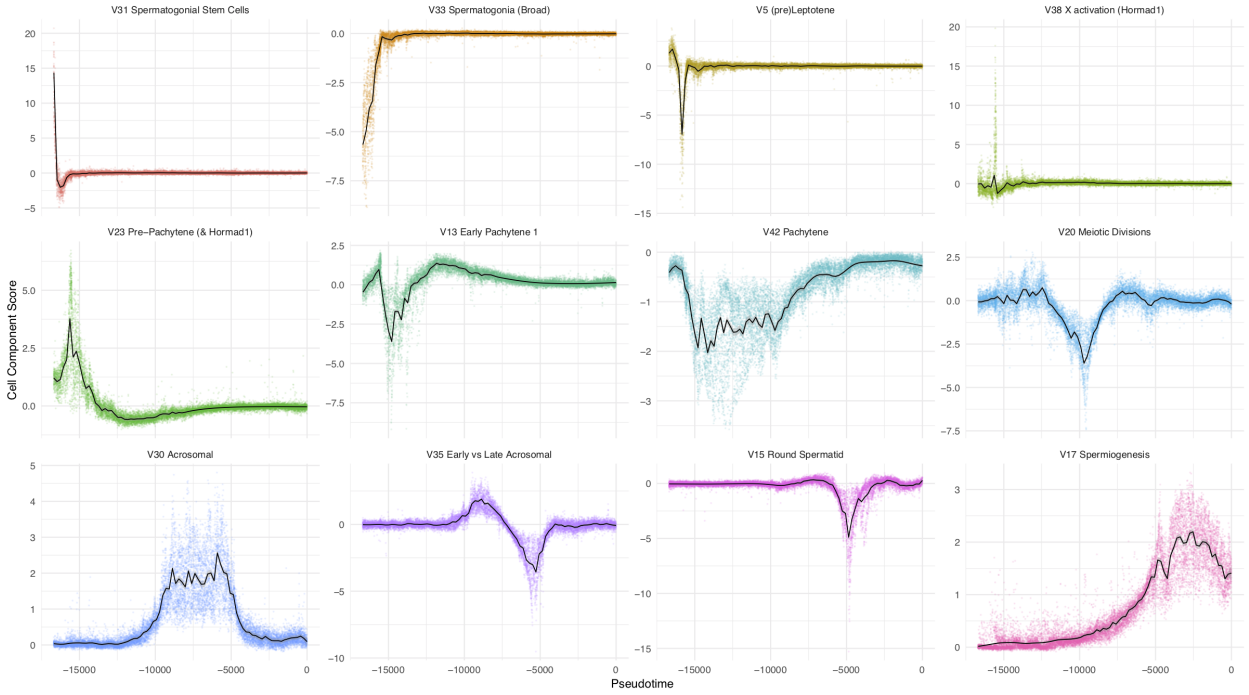


Figure 4 - Figure Supplement 2: Overview of Individual SDA Components.
Cell scores plotted as a function of pseudotime, for 12 representative SDA components.

In addition to identifying soft clusters and their markers, by multiplying the cell scores and gene loadings, SDA can impute very sparse, noisy, expression data. In principle, harnessing the correlation structure of gene coexpression across cells can improve predictions, overcoming the sparsity of the initial data. Indeed, our dataset has 93.8% zero values and a median of 1,312 UMI transcripts per cell. Nonetheless, SDA imputation is able to estimate expression of individual genes even when in many cells zero reads are observed (**Figure 5A**). It is not possible to determine the true expression vector for an individual cell, so we use cross-validation to test whether imputation improves expression estimates. Specifically, we assign each read to either a training or test set. We

predict gene expression based on the training set, using the SDA approach, or another approach (e.g. the dedicated single cell imputation method MAGIC [90]), and then evaluate our ability to rank gene expression using the test set (Methods). SDA imputation outperforms approaches using the raw data, for essentially all cells in the test data (**Figure 5B&C, Figure 5 Supplement 1C**). While providing the most sparse representation (**Figure 3C**), SDA still imputes equally well, compared to other matrix factorizations and to MAGIC [90](**Figure 5C, Figure 5 Supplement 1A)**. Further, when compared to NNMF, SDA provides additional biological insights for the same number of components (**Methods; Figure 5D, E,F & Figure 5 Supplement 1B**). In addition to obviating the need for further clustering and differential expression analyses, an advantage of using matrix factorization for imputation is the much smaller memory footprint required to store the results: on our dataset MAGIC data is 2.9Gb whereas the SDA matrices are just 18Mb (12.6Mb when loadings with PIP<0.5 are set to 0).

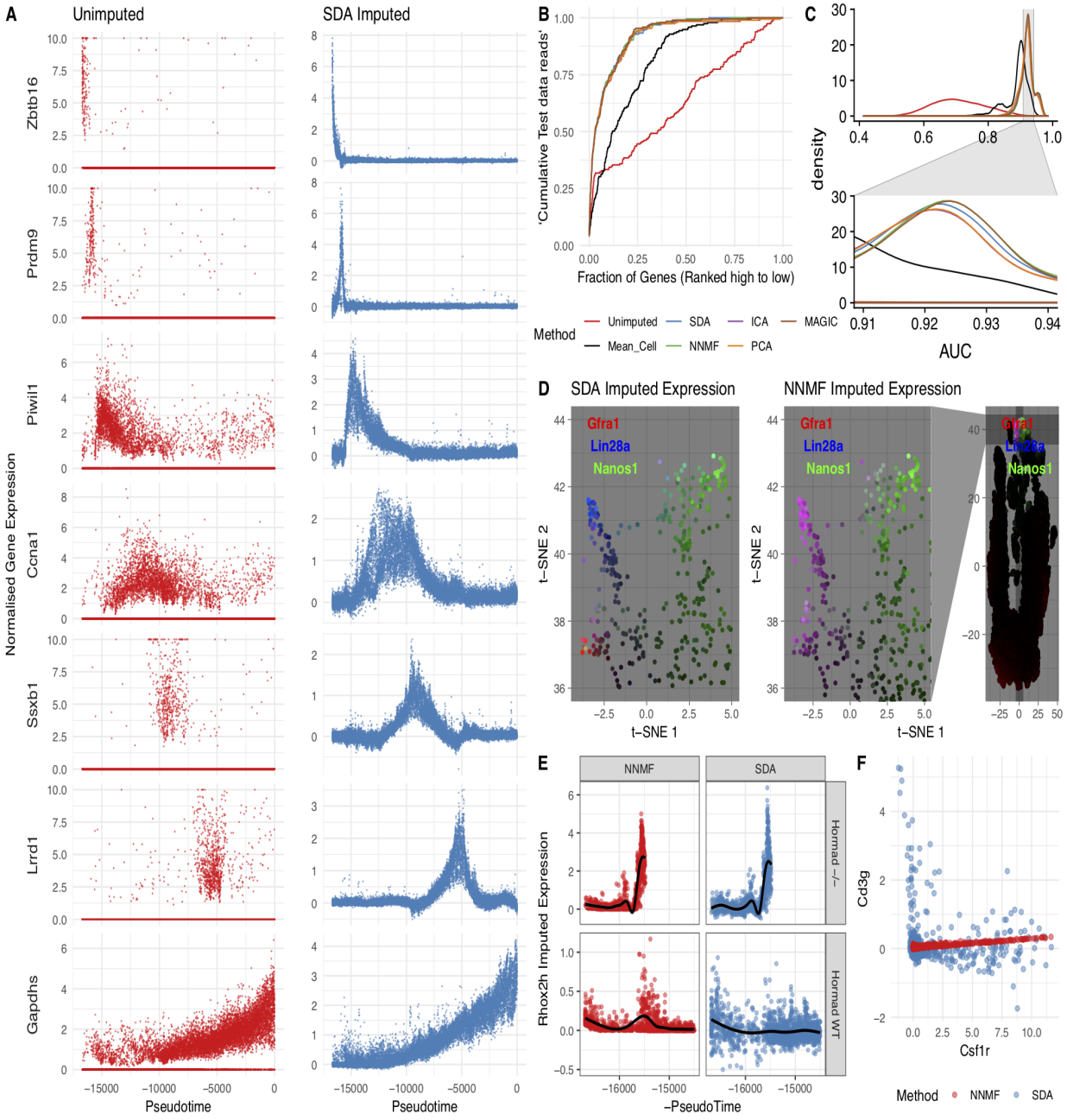


Figure 5: Evaluation of imputation using the SDA model:
(A) Here we illustrate the ability of SDA-based imputation (Methods) of gene expression values in single cells to improve the signal/noise ratio of expression, for 7 genes with strong developmental regulation. Note in the imputed expression "dropouts" at 0 are recovered and there is less outlying expression. **(B)**To test the utility of SDA-based imputation we created separate training/test data (**Methods**). From the training data we constructed seven predictors of gene expression in the test data for each cell ("Unimputed" using the training data directly, "Mean Cell" using the mean across all cells, matrix factorization approaches SDA, PCA, ICA, NNMF, and a dedicated imputation approach, MAGIC). We compared the ability of each predictor to rank the gene expression in the test data for each cell, quantified as the area under

the Rank Prediction Accuracy Curve (RPAC). Shown is an example RPAC for these predictors when applied to the test data for a single cell. **(C)**Comparison of AUCs (Area under the RPAC curve) for all cells using various methods (same color scheme as part B). **(D)** SDA produces multiple components for spermatogonia. Shown are zoomed in versions of the t-SNE projection (with full t-SNE for context): cells are colored by expression using a three channel ternary color scheme with the amount of blue, green, red representing the respective expression levels of Lin28a, Nanos1, and Gfra1. By assigning only one component for undifferentiated spermatogonia, NMF predicts Gfra1and Lin28aare expressed in the same cells resulting in a pink hue (See also **Figure 5 Supplement 1B,**no correlation for SDA component 50 Gfra1 Stem Cells). For selection of component see **Methods**. **(E)**Imputed expression of X chromosomal gene Rhox2hfrom either the SDA or NNMF decomposition, split into cells we know to be either WT or Hormad -/- genotype. NNMF predicts a peak in Rhox2h expression even in the WT cells, in which X chromosome activation due to Hormad1KO does not occur. (**F**) NNMF does not assign separate components for the innate and adaptive immune cells (See also **Figure 5 Supplement 1B**, no correlation for the SDA component 3 Lymphocytes). NNMF does not predict high expression of the adaptive immune cell marker Cd3g(T-cell surface glycoprotein CD3 gamma chain), and when it predicts any expression it increases linearly with the innate immune cell marker Csf1r(Macrophage Colony-Stimulating Factor 1 Receptor, or Cd115). SDA on the other hand correctly predicts that Cd3g and Csf1 rare not coexpressed in the same cells.

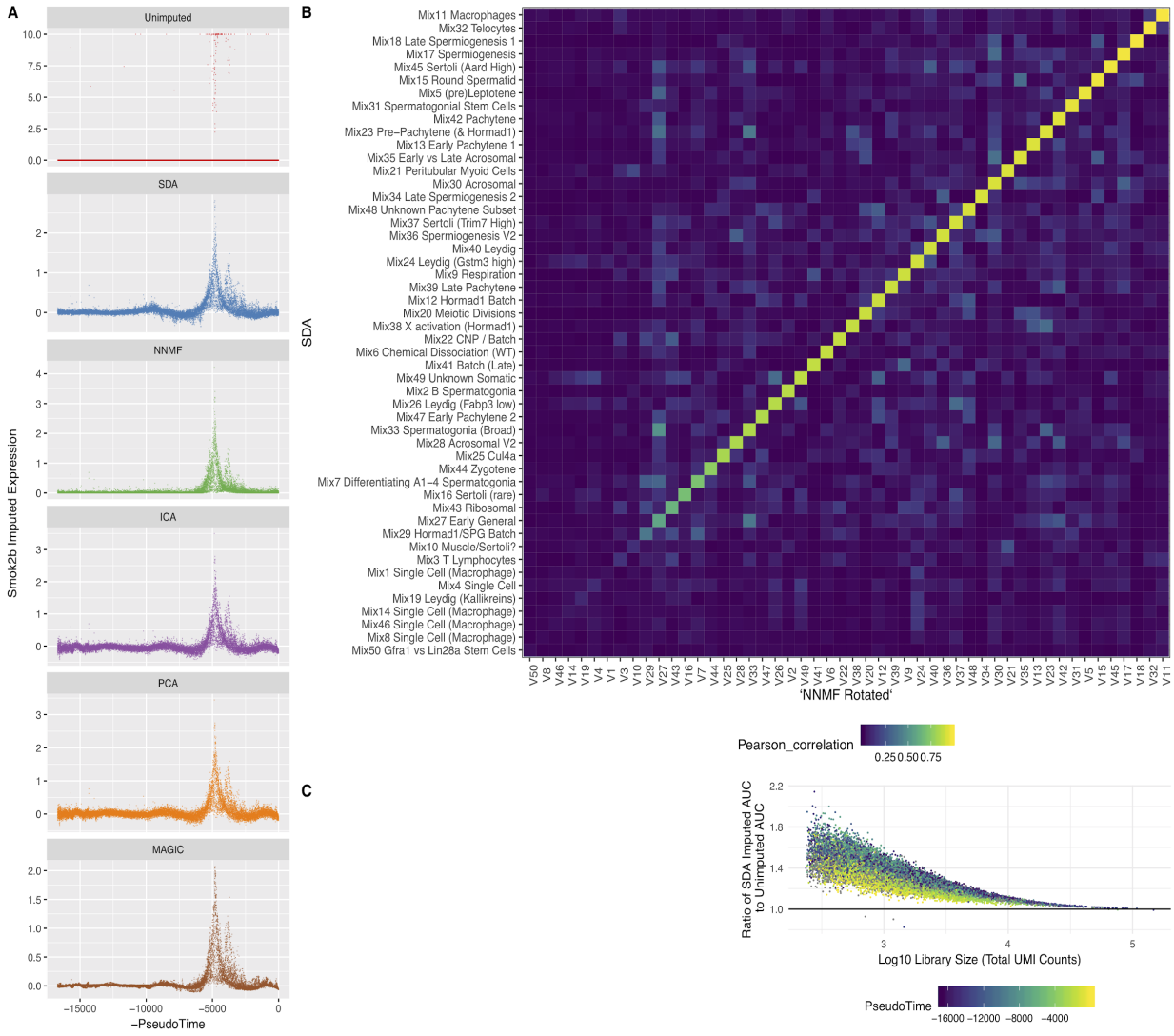


Figure 5 - Figure Supplement 1: Imputation from SDA and Other Matrix Factorization Methods

(A) Imputed expression of an example gene (Smok2b) for different methods, to illustrate the similar predictions as shown in **Figure 5B and C**. (B) Overall, NNMF infers similar components to SDA. The heatmap shows Pearson correlations between different pairs of gene loading vectors from SDA and NNMF (with procrustes rotation applied, **Methods**). (C) The fold improvement in AUC when comparing SDA imputation to the unimputed data, plotted as a function of cell library size. The gain in accuracy from SDA imputation is inversely correlated with library size i.e. the cells with low UMI count have most to gain.

Overall, of 50 components, 6 represent batch effects, 5 are components with only a single cell, 13 are observed only in somatic cell types, 23 only in germ cells, and 3 components load on

both somatic and germ cells (**Figure 3 - figure supplement 1**). Within somatic cell components, we observe components corresponding to Sertoli cells (n=4), Leydig cells (4), macrophages (1), T lymphocytes (1), telocytes (1), peritubular myoid cells (1) as well as an interesting component that seems expressed in all interstitial cells (1). Among germ cell-specific components, we observe components corresponding to processes active in spermatogonia (5), preleptotene spermatocytes (1), leptotene/zygotene (2), pachytene (5), diplotene (1), and spermiogenesis (7). Thus, we find multiple sub-components within existing recognized meiotic stages, adding considerable resolution relative to bulk-sequencing approaches. For some analyses below, we considered positively and negatively weighted genes within a component separately, in case these represent different modes of regulation, within the same groups of cells. We provide a web application to enable interactive exploration of gene expression and components at http://www.stats.ox.ac.uk/~wells/testisAtlas.html.

3.3.5. Components Reflect Known Biology But Also Highlight Sets of Genes With Mysterious Purpose

Five components correspond to processes in spermatogonia. Component 31 represents undifferentiated spermatogonia expressing *Zbtb16*(aka *Plzf*) [91]and *Foxo1*[92], while component 51 splits these cells into two subpopulations one expressing , *Gfra1*[93]and *Glis3*[94], and the other *Nanos3*[95],*Lin28a* [96]and *Foxf1*(**Figure 5D**). Component 7 likely represents $A_{1\text{-}4}$ spermatogonia expressing *Glis2*, *Nanos1*, *Kit,* and *Stra8*. Component 2 includes *Ctcfl*, *Pou4f1*, and *Esx1*- likely representing intermediate and type B spermatogonia and component 33 is a broader spermatogonial component enriched in genes involved in spermatogonial differentiation (**Table 3**).

During meiosis there is an extended prophase I (lasting 14 days in mice), which is itself divided into a number of stages: Leptotene, Zygotene, Pachytene, and Diplotene [97]. During these stages homologous chromosomes must pair to enable genetic recombination and balanced segregation during meiotic divisions. In the earlier stages homologous chromosomes begin to associate aided by meiosis-specific cohesin and tethering of the telomeres to the nuclear envelope [98,99]. Several hundred programmed double-strand breaks (DSBs) are then induced by *Spo11* at sites marked by *Prdm9*[100–103]. These DSBs are then resected to form single stranded DNA enabling homology search and repair which occurs in the context of a proteinaceous scaffold named the synaptonemal complex [104].

As an illustrative example, we focus on component 5, marking Leptotene. In this component, many of the genes required for these coordinated processes have high (top 500) loadings, including *Prdm9* itself; components of the meiotic cohesin complex *Rad21l*, *Smc1b*, *Smc3*, *Stag3*and *Esco2*[105]; components of the telomere tethering complex *Terb1*, *Terb2*, *Spdya,*and *Sun1*[106–108];genes involved in creating DSBs *Mei1*,*Ccdc36*(*Iho1*), *Spo11* partner *Top6bl*(*Gm960*), and regulator *Atm*[109–113]; proteins required for the creation and processing of the ssDNA intermediates and their regulators: *Mcm8*, *Dmc1*, *Rad51*, *Rad51ap2,Atr*, *Brca2*, *Tex15*, *Meilb2*(*Hsf2bp*), *Meiob*, and *Spata22*[114–125]; class I crossover (ZMM group) proteins *Shoc1*(*Zip2* orthologue), *Tex11*(*Zip4* orthologue), *Msh5*,*Hfm1*(*Mer3* orthologue) and regulator *Brip1*(*FancJ*) [126–130]; as well as components of the synaptonemal complex *Sycp1*, *Sycp2*, *Sycp3*, *Syce2*, *Syce3*, *Tex12*, and *Six6os1*(*4930447C04Rik*) [131,132]. (**Figure 3D-F**; **Table 3**).

Strikingly, this same component is highly enriched for GWAS hits of recombination rate in humans[133]. Of the 24 significant GWAS loci identified with confidently associated causal genes, more than half (13) rank within the top 300 genes of this component, and almost all (20) rank

within the top 1300 genes ($p = 5.2x10^{-18}$, OR = 77.8 and $p = 2.4x10^{-20}$, OR = 70.1 respectively by Fisher's exact test [FET], **Figure 3F**). Another hit, *Msh4*, is not ranked highly in this component (2734th out of 19262), however, its product is known to function as a heterodimer with the product of *Msh5*, which ranks 34th [130]. This highlights one of the advantages of single cell RNA-seq compared to GWAS for target discovery in that it does not rely on the presence of (perhaps rare, small effect) genetic variants. In addition, it directly provides a list of genes rather than SNPs affecting unknown causal genes. For example a previous GWAS [134]had identified a SNP in the intron of *Ccdc43*, however our expression data strongly suggested the adjacent gene *Meioc* (*aka C17orf104*) as the causal gene (ranked 183rd vs 13,651st in component 5), providing additional evidence relative to reports that *Meioc* is responsible for maintaining an extended meiotic prophase [135,136]. Indeed, the lead SNP in this region in a more recent GWAS is in the promoter of Meioc [133]. The strong enrichment of genes involved in recombination in this component suggests other highly ranked genes of unknown function could also play key roles in this process. During the preparation of this manuscript, two such genes were identified: *Ankrd31*(ranked 102nd) plays a role in controlling the number, timing, and location of double strand breaks in meiosis [137,138],while *Hsf2bp*(now *Meilb2*, ranked 194th) was found to be a master regulator of meiotic recombinases [116].

One striking candidate gene is *Zcwpw1*, which ranks 3rd, after *Prdm9*. This gene does not have a known function, but contains two protein domains: CW and PWWP, known to bind H3K4me3 and H3K36me3 respectively [139,140]. PRDM9 deposits both H3K4me3 and H3K36me3 at sites it binds [141], and this methyltransferase activity is essential for its role in double strand break positioning (Diagouraga et al., 2018), suggesting these marks may be recognized by downstream protein(s). An obvious hypothesis is that ZCWPW1 might co-localize to recombination hotspots,

by binding the histone modifications deposited by PRDM9. Further work will be required to test this, and the potential role of ZCWPW1 in meiotic recombination.

Prior to single cell studies such as our own, previous approaches to germ cell transcriptional profiling provided a single, static summary of pachytene expression from bulk sequencing of purified cells [142,143]. Here, we are able to decompose pachytene gene regulation into 5 components (13, 39, 42, 47, and 48). Although the cell loadings for these components overlap in pseudotime, they differ dramatically in their dynamics (**Figure 4A&B**). For instance, component 13 and 47 loadings appear to fluctuate, from positive, to negative, to positive again, while component 42 loading is constantly negative when active. The genes with strong loadings within expression components do not necessarily associate with a single, coherent functional process, nor even a set of transcripts that are all translated at the same point in spermatogenesis. Instead, components 13, 39, 42 and 48 each appear to involve both a substantial number of genes required for meiosis, but also a second set of genes needed for some postmeiotic processes, including genes involved in sperm tail formation (**Table 3**).

The early pachytene components 13 and 47 are enriched for genes involved in the meiotic cell cycle (e.g. *Ccna1*, *Cdk1*), chromosome pairing and segregation (e.g. *Sycp3*, *Dmc1*, *Hormad1*), nuclear division (e.g. *Cenpe*, *Plk1*), and piRNA processing (e.g. *Tdrd1*, *Tdrd5, Tdrd9, Piwil1* and *Piwil2*). The next component in the temporal sequence, 48, is restricted to a small cluster of cells in t-SNE space and enriched for many genes involved in axoneme/cilia assembly (multiple members of the *Cfap* family and dynein genes) and a smaller number of genes involved in microtubule/spindle formation (e.g. *Dcdc2b*, *Ccdc88a*, *Knl1*) and RNA splicing (e.g. *Srrm2*, *Tra2a*, *Srek1*). Components 42 and 39 (pachytene/late pachytene) are enriched for distinct genes, enriched

for similar biological functions - such as meiotic cell cycle, cilium assembly, piRNA processing, and translational suppression.

The pachytene components have a striking lack of genes loading on the X or Y chromosome (**Figure 6E**), due to meiotic sex chromosome inactivation (MSCI), which is part of a broader mechanism silencing unsynapsed chromatin (MSUC) [144,145].MSCI is an evolutionarily conserved phenomenon essential for proper spermatogenesis in mammals. As previously reported [82–84]we observe MSCI from the start of pachytene (**Figure 6A & Figure 4 - figure supplement 3D**). Although previous bulk RNA-seq studies suggested that some genes escape MSCI [142,143], we were unable to confidently identify any genes escaping MSCI. A small number of sex-chromosome transcripts identified in pachytene cells were observed; however, these genes were highly expressed in neighboring Sertoli cells, suggesting low-level contamination as the most likely explanation. Moreover, our data indicate that previously identified "escapees" are actually expressed after MSCI, yet fully silenced within MSCI (**Figure 6C & Figure 6 - figure supplement 1B**).

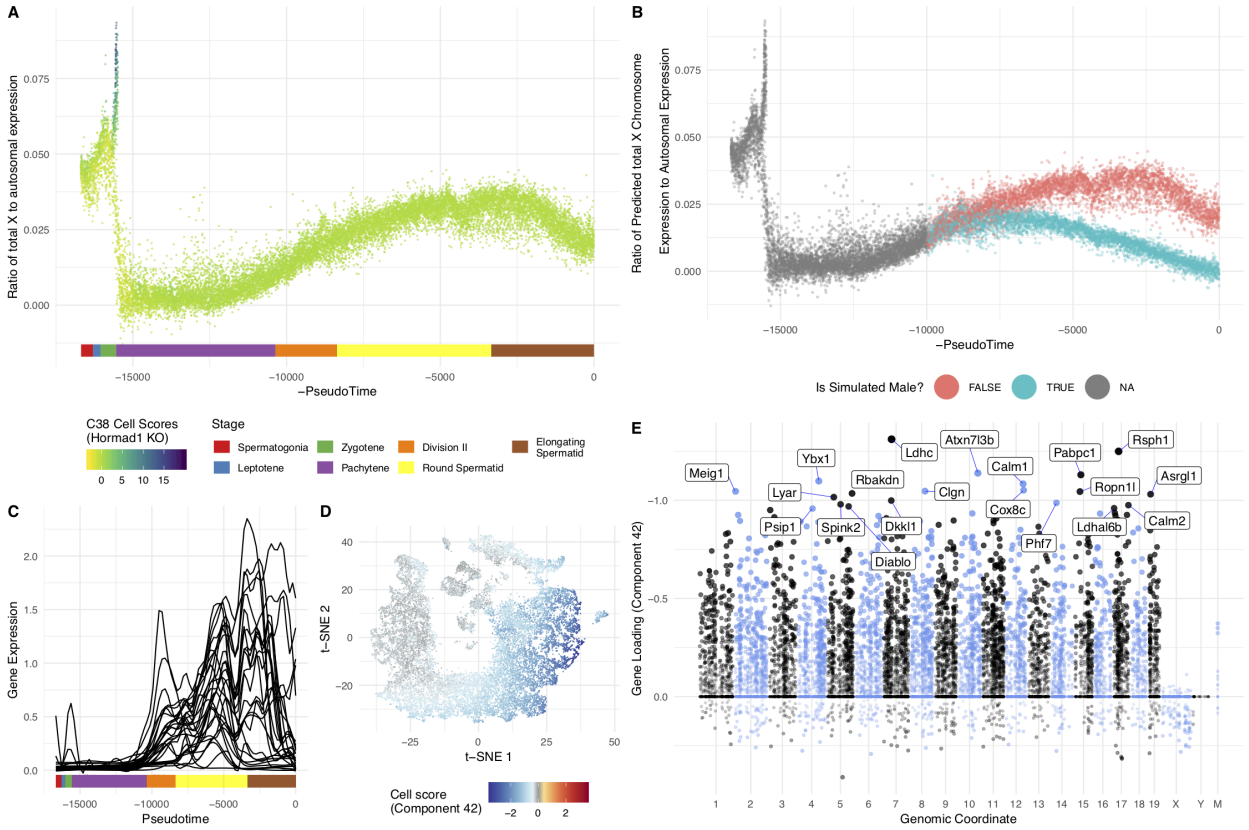


Figure 6: Insights into sex chromosome biology from SDA.
(A) Pseudotime analysis provides quantitative, high-resolution insights into meiotic sex chromosome inactivation (MSCI). The sum of imputed expression for all genes on the X chromosome divided by that of the autosomes (y-axis) drops to almost 0, showing near-complete MSCI before gradually partially recovering. A similar profile is observed for genes on the Y chromosome (Figure 6 - figure supplement 1A). (B) We do not observe that haploid cells obviously split into two populations due to lack of sex chromosome transcript sharing, in part A. Here were simulate what we might expect to see if there was indeed a lack of sharing (Methods). (C) No evidence supporting prior report of genes escaping MSCI. Smoothed expression values (unimputed, gam smoothing with formula "y ~ s(x, bs = ad)") are shown for each gene reported to escape MSCI 142excepting H2al1e, H2al1c, and Gm10096which were below our dataset's expression detection threshold. Expression profiles for individual genes are separated in Figure 6 - figure supplement 1B. (D) Component 42 (Pachytene) cell scores in t-SNE space. (E) Component 42 gene loadings. This component represents genes active during the pachytene stage of meiosis; note the striking lack of sex chromosome gene loadings, due to MSCI.

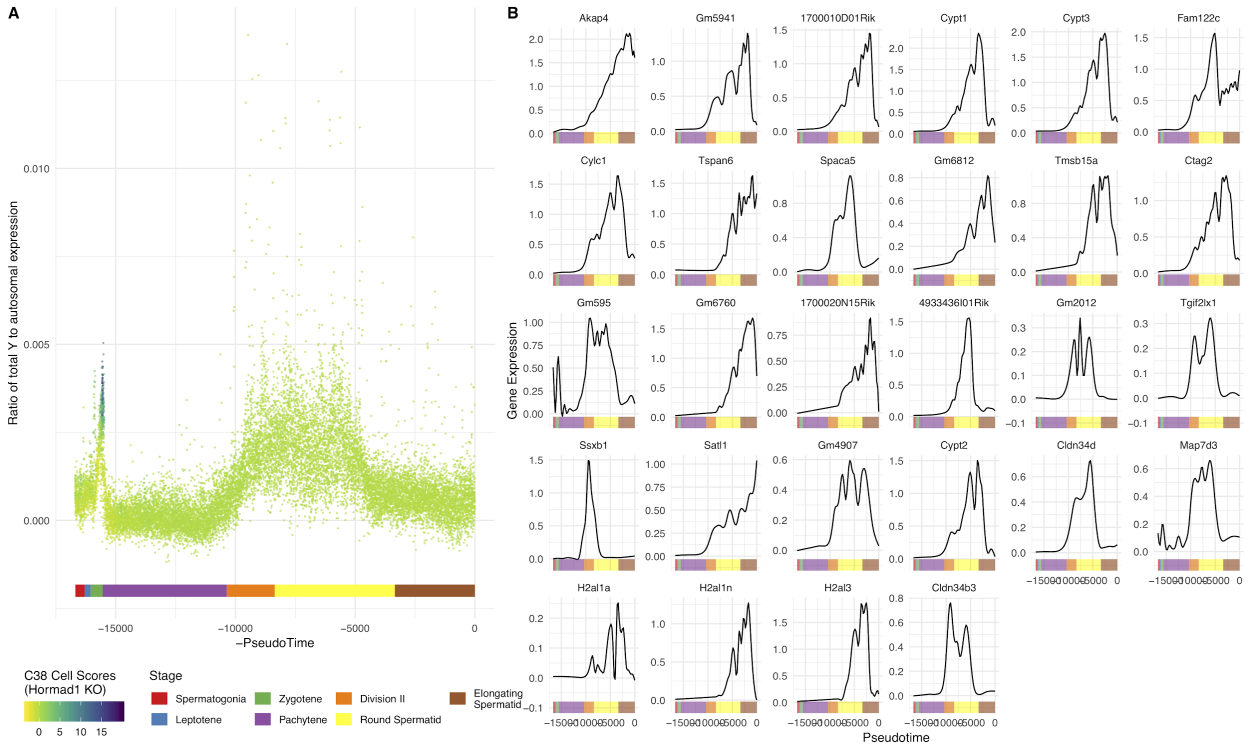


Figure 6 - Figure Supplement 1: Single-gene analysis of MSCI.
(A) As for Figure 6A, but Y chromosome instead of X. (B) As for Figure 6C, but each gene is shown individually.

In addition to MSCI there is the potential for lack of sex chromosome transcripts later in meiosis, due to the fact that post-meiotic cells have haploid genomes, meaning they have either an X or a Y chromosome but not both. However, cytokinesis does not fully complete in spermatogenesis resulting in synchronized chains of hundreds of cells, connected by µm-wide cytoplasmic bridges through which mRNA (or perhaps even mitochondria) could be shared [146]. The extent to which mRNA sharing occurs is unknown, but it is a property of interest to evolutionary biology as most models predict a strong fitness benefit to fathers who can mask haploid selection in their gametes [147]. Here, we find that, with respect to sex chromosome transcription, the genetically haploid cells are predominantly phenotypically diploid (**Figure 6A & B, and Figure 6 - figure supplement 1A**) suggesting that cytoplasmic mRNA is efficiently

shared, consistent with studies of individual genes [2]and recent scRNA-seq reports [82,83]. However, there remains a possibility that some genes are not shared, such as has been observed for autosomal genes in a mutant heterozygous context: the t-complex responder mutant ($Smok^{Tcr}$) which functions as an antidote in the poison-antidote meiotic drive system of the t-complex [148]and Spam1 which causes transmission ratio distortion in Robertsonian (Rb) translocation-bearing mice [149].

Component 20 is particularly interesting. Genes in this component are likely to be functional during meiotic divisions and perhaps afterwards. It contains a number of genes known to be expressed in diplotene and/or key regulators of cell division, in addition to the *Ssx* family of genes (discussed further below) and also shows very strong enrichment of genes characterized by the presence of a DUF622 domain (18 in the top 88 genes) (**Table 3**). This gene family is rodent-specific and arose from duplication of the gene *Dlg5* [150].It has previously been shown that many, autosomal, DUF622 genes experience similar epigenetic changes to the sex chromosomes during spermatogenesis [151]. Another component (9) is most active at a similar time to 20 and is very highly enriched for genes of the electron transport chain ($p = 7.4x10^{-53}$, OR = 104, FET) (**Figure 4 - figure supplement 4E&F**).

We identified 7 post-meiotic components characterizing wildtype biology. Round spermatid component 30 contains many genes associated with the acrosome, an organelle which forms a nuclear cap containing hydrolytic enzymes used in fertilization [152](**Supplementary File 3**). In addition, the gene *Lrrc34* has a high loading. We verified by immunofluorescence that the protein is indeed localized to the acrosome of round spermatids (**Figure 2B**). Component 35, which is essentially concurrent to component 30 in pseudotime, is the most mysterious of all components that we detected. Dozens of protein-coding genes in this component are highly enriched in testis expression but have no known function (**Table 3**). This component also harbors

a substantial number of genes with no apparent ortholog in humans. The existence of such a set of poorly characterized genes likely reflects the difficulty of studying post-meiotic male germ cells - these cells cannot be differentiated *in vitro*, contain numerous cell-type specific processes, and express many genes which are rapidly evolving.

The spermiogenesis components 17, 18 and 34 all contain many genes known to be expressed at the latest stages of spermatogenesis, before transcriptional arrest due to replacement of histones with protamines [153](**Table 3**). In addition, *Abhd5*(aka CGI-58), a protein previously detected in testis lipid droplets [154], has high loadings specifically in these late components (17 & 18) and we show by immunofluorescence that it serves as an excellent marker of the residual body (**Figure 2B**).

In addition to components for the germ cell transcriptional programs we identified components for at least 5 different somatic cell types: Sertoli, Leydig, macrophages, peritubular myoid cells, and T-lymphocytes. We also find a component representing an abundant somatic cell type expressing *Tcf21* but not *Acta2*, described by [83]as an unknown mesenchymal celltype, which we identify as telocytes based on co-expression of *Cd34* and *Pdgfra*[89,155]. Some components clearly mark multiple cell types that resolve separately in t-SNE space, while others mark groups of cells that may contain cryptic heterogeneity obscured by overlapping gene expression patterns (**Figure 4 - figure supplement 3F and Figure 3 - figure supplement 1**). We were also able to infer components for batch effects such as differences in sequencing machines and different individual mice (**Figure 4 - figure supplement 4A-D**).

3.3.6. Joint analysis of 5 mouse strains identifies pathology-related components

The flexibility of the SDA modeling framework allows the identification of sets of genes that show significant covariation in small numbers of cells. Thus, a joint analysis of mutant and wildtype cells using SDA could potentially decompose expression variation into separate technical effects, variation due to normal biological processes, and variation due to pathology, identifying mutant-specific components in the context of wildtype cells. We set out to evaluate the utility of single-cell sequencing to identify pathology in each mutant strain, combining results from both classical and SDA approaches.

Increased apoptosis is an important mechanism underlying many genetic forms of male infertility in mice. Apoptotic cells can be identified from single-cell RNA-seq data as having an excessive proportion of total transcriptome derived from mitochondrial genes [156]. Cells from *Mlh3*-/- and *Hormad1*-/- animals showed higher rates of apoptosis compared to wildtype, *Cul4a* and *Cnp* (2% vs 14.5%, **Figure 1 - figure supplement 1**). Pseudotime analysis provided an even finer level of resolution for staging the time of onset of developmental problems in each strain (**Figure 7A**). By performing joint pseudotime analysis on all strains simultaneously, it is in theory possible to fine map the timing of developmental defects. Our our pseudotime-ordered set of 16,950 germ cells spans the entire ~34.5 day [157]development process from Type A spermatogonia to mature spermatozoa, suggesting a mean difference in developmental age between pseudotime-adjacent cells of 3 minutes. Although further work is needed to clarify the mapping of pseudotime to real time, that mapping estimates the difference in the mean time of arrest of *Hormad1-/-* cells and *Mlh3-/-* cells to be 12 days. This difference is reflected in the SDA components as well; *Mlh3-/-*

animals possess cells that load on pachytene components 47, 42 and 39, while *Hormad1-/-*animals do not.

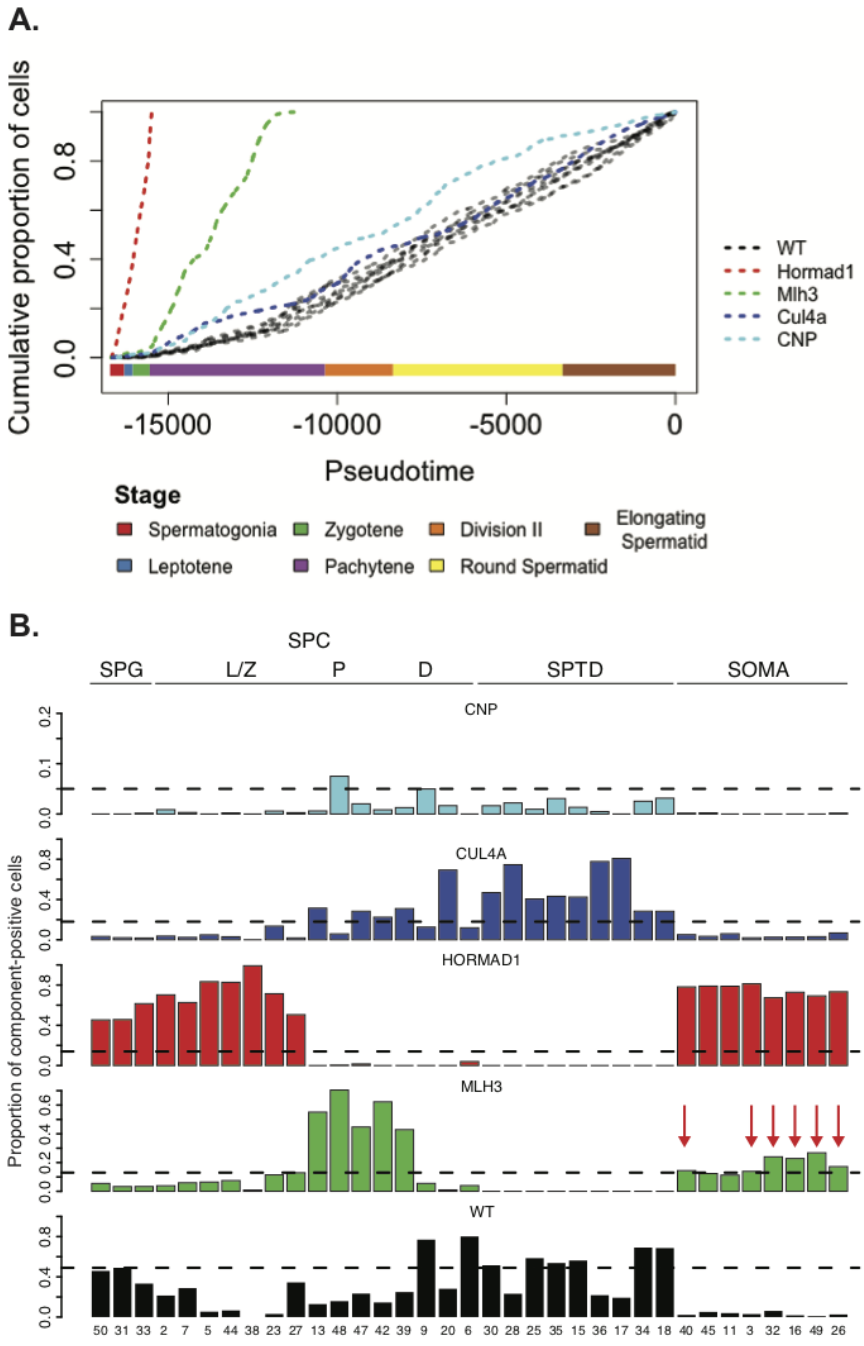
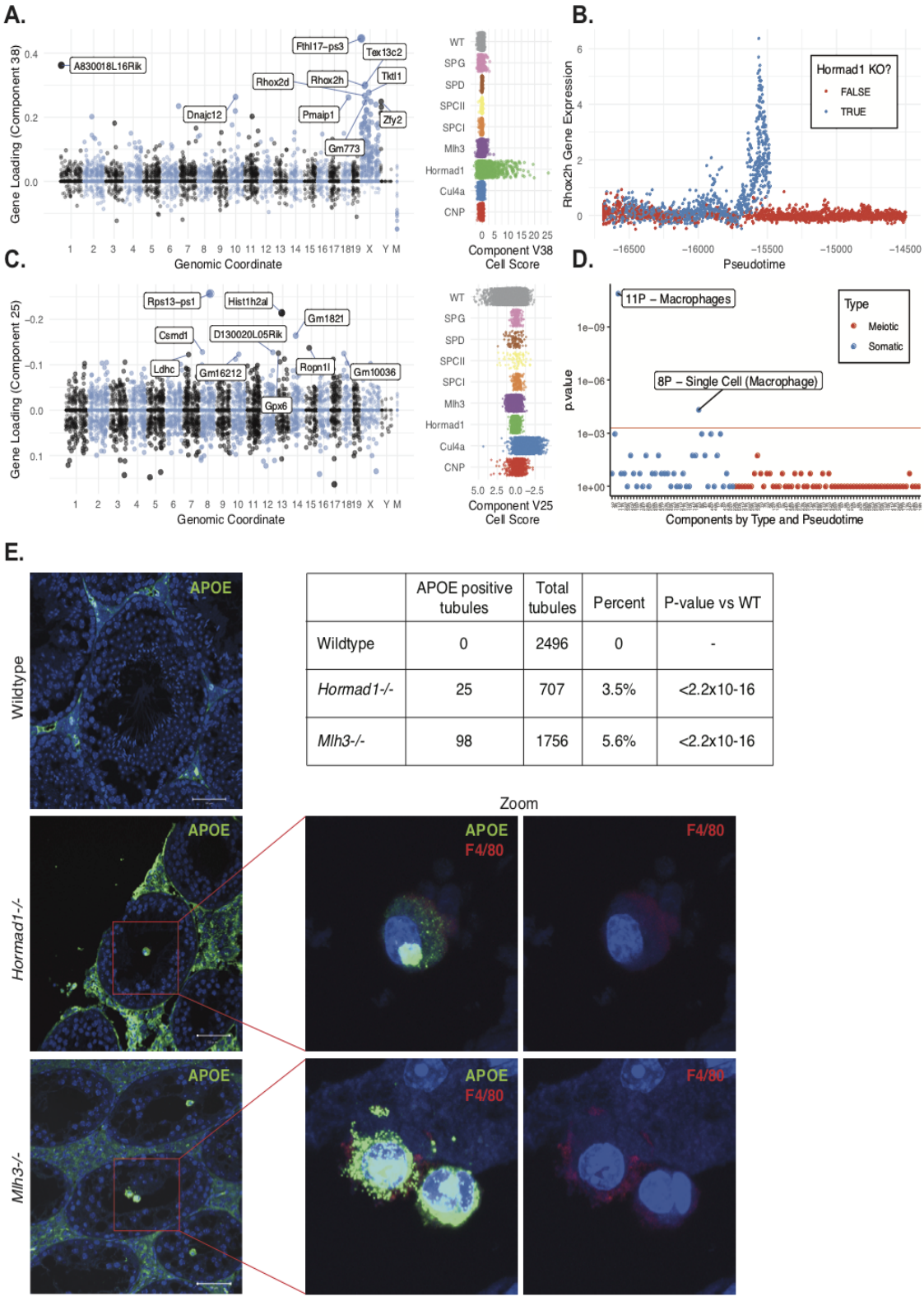


Figure 7: Characterization of mouse mutants with testicular phenotypes using pseudotime and SDA.
(**A**)The cumulative distribution of cells along pseudotime from each mouse strain. The data clearly indicate that Hormad1-/-cells arrest prior to Mlh3-/-cells in the pachytene stage of spermatogenesis, while Cul4a-/-and CNP mice show quantitative deviation from WT in the abundance of postmeiotic cells. (**B**) As a way to summarize the SDA analysis of each strain, we plot the proportion of cells with strong component loadings from each strain separately. If cells are randomly distributed across components then we would expect the fraction of cells from each mutant to be the proportion of total cells sequenced from that mutant (dashed horizontal lines). Instead there are clear enrichments of component loadings in particular mutants, providing a fingerprint of pathology for those strains. Arrows indicate six components (3, 16, 26, 32, 40, and 49) that were enriched for cells from both Mlh3-/- and Hormad1-/- strains, but no other strains. SDA components are sorted by developmental stage, as indicated by horizontal lines across the top of the panel. SPG = spermatogonial components; L/Z = leptotene/zygotene components; P = pachytene components; D = diplotene components; SPTD = components in spermiogenesis; SOMA = somatic cell components.

HORMAD1 is a meiosis specific protein that regulates chromosome recombination, synapsis, and segregation. HORMAD1 normally marks un-synapsed chromosomes (including sex chromosomes). While HORMAD1 is removed by TRIP13 on synapsis, it persists on asynapsed chromosomes, which then undergo MSUC, leading to MSCI for the sex chromosomes [158,159]. In *Hormad1-/-*spermatocytes, double-strand break formation and early recombination are disrupted as marked by the reduction of yH2AX, DMC1, and RAD51 foci [159]. Hard clustering analysis (**Figure 1F & G**) showed a deficit of post-pachytene *Hormad1-/-* germ cells, consistent with the expectation that *Hormad1-/-* mutant cells experience apoptosis during meiosis I due to pachytene checkpoint failure [160]. Along with this arrest phenotype, the *Hormad1-/-*leptotene/zygotene cells form a distinct cluster outside of the leptotene/zygotene cells of all other strains (Cluster 30, **Figure 1 - figure supplement 2 and 3**). A list of significant differentially expressed genes between the cluster 30 and neighboring cluster 32 included a number of sex chromosome genes **(Table 2**). Consistent with these observations, we found one SDA component (38) with much higher gene loadings on the sex chromosomes than autosomes (**Figure 8A, Figure 4 - Supplement 3C, Table 3**), and with cell loadings that are specific to *Hormad1*$^{-/-}$. We find that not only does *Hormad1*$^{-/-}$

fail to silence previously expressed sex-linked genes, but many previously *un*expressed sex-linked genes such as *Rhox2h*obtain high expression (**Figure 8B**). Interestingly, there are also multiple autosomal genes with high loadings. This may be due to ectopic expression of sex-linked transcription factors; for example, *Zfy1*and *Zfy2* were previously shown to cause pachytene arrest when misexpressed [161]. We find a very strong association between genes in this component and genes overexpressed in mice which have mutations in either *Hormad1*or *Trip13*($p = 2.2x10^{-39}$, OR = 184 and $p = 1.3x10^{-157}$, OR = 115, respectively by FET) [162]**; Figure 4 - figure supplement 3A&B**.



	APOE positive tubules	Total tubules	Percent	P-value vs WT
Wildtype	0	2496	0	-
Hormad1-/-	25	707	3.5%	<2.2x10-16
Mlh3-/-	98	1756	5.6%	<2.2x10-16

Figure 8: Dissection of strain-specific pathology.

(**A**) SDA component 38 is comprised largely of genes on the X chromosome, with a gene loading direction that indicates failure of X inactivation. As illustrated by the cell scores (loadings) for this component, it is restricted to Hormad1-/-cells. (**B**) Pseudotime analysis indicates that Hormad1-/-cells diverge developmentally from all other strains around leptotene/zygotene. In this illustration, the X-linked gene Rhox2his shown to have low or no expression in all cells prior to meiosis, and then rapidly increased expression specifically in Hormad1-/-cells until this lineage arrests. (**C**) Component 25 is the component most strongly enriched for Cul4a-/- cells. (**D**) We identified 6 components with shared enrichment for both Mlh3-/-and Hormad1-/-cells; these components contained genes with numerous significant GO associations related to Alzheimer's disease (AD) pathology (main text, **Figure 8B**). For each SDA component, we tested for association between known AD genes and genes with either positive (P) or negative (N) loadings on that component. AD genes are highly enriched for expression in component 11, corresponding to macrophages. (**E**) Further investigation of protein expression of AD genes revealed APOE+ (green) cells within the tubules of Mlh3-/-and Hormad1-/-but not WT. These cells showed nuclear morphology different from native germ cells or Sertoli cells, and stain positive for the macrophage marker F4/80. The inset table summarizes raw data on the frequency of APOE+ tubules obtained by microscopy. The frequency of APOE+ tubules is more common in each mutant strain when compared to WT by Fisher's exact test. Scale bar = 50μm.

CUL4A is a major component of the E3 ubiquitin ligase complex called CRL4 which is known to regulate cell cycle, DNA replication, DNA repair, and chromatin remodeling [163]. Studies on *Cul4a-/-*mice noted that some spermatocytes arrest at the pachytene stage of meiosis I induced by the pachytene checkpoint, whereas remaining spermatocytes complete meiosis but the resulting spermatozoa present oligoasthenospermia and severe malformations [164]. The molecular basis of observed abnormal phenotypes in spermatozoa remains unclear. We identified a single SDA component (25) that was highly specific to *Cul4a-/-*cells (**Figure 8C and Table 3**). This component corresponds to dozens of genes that are overexpressed in *Cul4a-/-*mutants when compared to all other strains, with GO enrichments related to spermatid development, motility and capacitation. These findings are consistent with the observed phenotype of *Cul4a-/-* mice and provide new leads to investigate mechanisms of pathology.

MLH3 is an essential protein required for crossover formation in early meiosis and for binding of MLH1 to meiotic chromosomes. Studies on *Mlh3-/-*testes have shown depletion of

spermatocytes and some spermatogonia due to apoptosis in diplonema induced by a reduction of chiasmata and a loss of recombination nodules [165]. Interestingly, in contrast to *Hormad1-/-*, we found no obvious transcriptional phenotype in *Mlh3-/-*cells either by SDA analysis or by comparison of expression levels between hard-clustered wildtype and mutant cells (other than differential expression of *Mlh3*). Instead, *Mlh3-/-*spermatocytes might simply trigger apoptosis through existing checkpoint protein machinery assembled earlier in development. Using the simple pseudotime analysis described above, we can estimate that if a transcriptional response was triggered, it might last less than ~33 minutes, for it to be missed in our sample of cells (**Figure 7A**). Similarly, the cells from *Cnp* mutant mice did not form distinct clusters, nor did they show SDA component loadings distinct from wildtype cells. Although the presence of multinucleated giant cells, hypocellular seminiferous tubules and infertile phenotype in these mice point to a serious defect in spermatogenesis, it seems difficult to determine which stages are affected using single-cell expression data. One possible explanation of missing important biological signals may be that *Cnp* mice presents a partial arrest phenotype which masked the developmental abnormalities. Another possible explanation is that droplet-based sequencing library preparation may undersample the cells with aberrant transcriptional signatures, e.g. due to failure of oil droplets to encapsulate the giant cells.

3.3.7. Invasion of macrophages into the seminiferous tubules is a convergent phenotype of meiotic arrest mutants

Despite the differences in cell composition or component loadings among mutant strains, we identified 6 somatic components (3, 16, 49, 40, 26, and 32) showing a specific enrichment for *Mlh3-/-*and *Hormad1-/-*cell loadings when compared to all other strains (**Figure 7B**). Hypothesis-free GO enrichment analysis of these components (**Methods**) revealed a recurrence of amyloid

related GO terms with qvalue < 0.01, with these terms being the highest enriched term in three components (26N, 49N, 16N). Excessive production of amyloid-beta, a primary cause of Alzheimer's disease, was not previously reported in these mutants, and the possible physiological role of such production is unclear. We tested multiple antibodies to human amyloid-beta that failed to work on our tissue. To further evaluate the expression of Alzheimer's disease (AD)-related genes across all 5 mouse strains, we tested individual SDA components for enrichment of expression of AD risk genes identified in a recent GWAS, identifying component 11 (macrophages) as specifically and strongly enriched ($p<10^{-12}$, **Figure 8D** and **Figure 4 - figure supplement 3E**). Immunofluorescence staining for the protein product of one well-studied AD gene, *Apoe*, in wildtype animals showed low levels of specific staining, confined to the interstitial space (**Figure 8E**). Both *Mlh3-/-*and *Hormad1-/-* displayed interstitial cell with more intense staining of APOE, as well as a greater abundance of APOE+ cells. More surprisingly, we also found a rare population of APOE+ cells within the tubules of *Mlh3-/-*and *Hormad1-/-*, that was never observed in wildtype. We screened 4,959 tubule cross-sections to establish more precise estimates of APOE+ cell frequency in these three lines (**Methods**). When compared to the frequency in wildtype tubules (0/2496 tubules), we see higher frequencies of intratubular APOE+ cells in *Mlh3-/-*(25/707 tubules, 3.5%, $p< 2.2 \times 10^{-16}$) and*Hormad1-/-*(98/1756 tubules, 5.6%, $p< 2.2 \times 10^{-16}$). These APOE+ cells displayed a nuclear staining and morphology that are distinct from normal germ cells and Sertoli cells and appeared more similar to APOE+ cells outside of the tubules. These APOE+ intratubular cells stained for F4/80, a well-established macrophage antigen, perhaps surprisingly, given that it suggests that in these mutants, immune cells can transit the blood-testis barrier and enter an area typically regarded as immune-privileged. Intratubular macrophages have rarely been described previously, again nearly always in the context of testicular defects [166–168]. Co-staining of F4/80

with an antibody for activated CASPASE-3, a marker of apoptosis, failed to identify any double positive cells, excluding the possibility that intratubular F4/80 protein expression was somehow an artifact of an apoptotic cell population. The mechanisms by which macrophages transit the blood-testis barrier, and the corresponding cues for migration, await further investigation.

3.4 Discussion

The extensive cellular heterogeneity of the testis has limited the application of genome technology to the study of its gene regulation and pathology. Here, we described how the SDA analysis framework can be applied to single-cell RNA-sequencing data of the testis to overcome the challenge of heterogeneity by summarizing gene expression variation into components that reflect technical artifacts, cell types, and physiological processes. Rather than clustering groups of *cells*, SDA identifies components comprising groups of *genes* that covary in expression and represents a single cell transcriptome as a sum of such components. This revealed previously uncharacterized complexity, with multiple different components even within recognized meiotic stages such as pachytene. This finer granularity suggests new biological interactions, for example the extremely high expression of *Zcwpw1*, a reader of specific histone modifications, within the same component as *Prdm9*, which induces identical modifications. We also identified components, both meiotic and non-meiotic, corresponding to interpretable pathology and specific to one or more mutant strains.

Other matrix factorization methods have been previously applied to soft cluster high dimensional gene expression data for example ICA, PCA [83,169], Bayesian Factor Analysis [170]and Non-Negative Matrix Factorization (NNMF) [171,172]which naturally has a degree of sparsity in both the cell scores and gene loadings due to the positivity constraint. More recently these methods have also been applied to single cell RNAseq data [173–178],reviewed in [179]. Here, we have reported

some comparisons between SDA and these standard methods. NNMF is often motivated by the positive nature of the original data, in addition to potentially increased interpretability for purely positively additive components. However, we note that latent factors, such as those utilized by SDA, which allow negative loadings have the potential to better capture transcriptional repression. In our tests, SDA retains similar imputation performance to NNMF, while providing a more compact (in terms of sparsity) representation of the data – aiding our interpretation of components found. Beyond matrix factorization, there are other frameworks with similar goals that have been applied successfully to single cell data. One set of methods are those based on neural networks, such as self-organizing maps [180]and deep-network autoencoders (DCA) [181]. DCA, much like tSNE, creates a nonlinear embedding of the high dimensional data resulting in a lower dimensional set of scores for each cell. This approach does not, however, provide the equivalent to gene loadings and so one would have to do additional differential expression analysis on a hard clustering of the latent embeddings in order to find genes associated with the latent dimensions. To assist comparison of SDA to other methods with overlapping objectives, we have summarized resource usage of SDA across a variety of run parameters, and input data sizes (**Table 4**).

Run name	Data Type	# Cells	# UMIs	# Genes	# Components	# Iterations	CPU time*	Memory Usage*
Mouse	DropSeq	20322	50278390	19262	25	10000	36 hours	9.2GB
Mouse	DropSeq	20322	50278390	19262	50	10000	59 hours	9.3GB
Mouse	DropSeq	20322	50278390	19262	70	10000	72 hours	9.3GB
Human	10X Genomics	15521	182942716	18589	50	10000	48 hours	7.0GB
Human	10X Genomics	15521	182942716	18589	70	10000	57 hours	7.0GB
Human	10X Genomics	15521	182942716	18589	100	10000	67 hours	7.1GB

**average of 5 SDA run results on a variety of 2.20 GHz-3.00 GHz processors*
***These SDA runs were not performed with parallel computation*

Table 4. Summary of SDA runtime and memory usage for example datasets.

Finally, we note that gene expression components (for example those identified by SDA) represent an attractive way to build a dictionary of pathology of the testis. The construction of new component models using a larger panel of mutants with known pathologies will accelerate the

interpretation of idiopathic mutants, and, ultimately, could provide a framework for a much more advanced diagnosis of human infertility than is currently in practice.

3.5 Materials and Methods

Mice

All animal experiments were performed in compliance with the regulations of the Animal Studies Committee at Washington University in St. Louis under protocol #20160089. Mice were housed in a barrier facility under standard housing conditions with *ad libitum* access to food and water and a 12hr:12hr light/dark cycle. All single-cell RNA sequencing experiments were carried out with sexually mature animals (ages of mice in this paper vary from 11- 38 weeks) except for *Pou5f1-EGFP* transgenic animal testes which were collected at post-natal age (P) 7. For specific age of mouse at the time of testes collection for different batches, please refer to **Table 1.** Samples for histological studies were also collected at the time of testes collection for single-cell RNA sequencing. The mouse lines used in this paper are the following:

1. C57BL/6J male mice were used for Hoechst-FACS and total testis single-cell RNA sequencing experiments.
2. B6;CBA-Tg(*Pou5f1*-EGFP)2Mnn/J reporter mice were used for enriching and isolating spermatogonia type A cells. Testes from five mice at post-natal age P7 were pooled to generate single-cell suspension and FACS sorted for GFP positive cells, followed by Drop-seq.
3. B6.129-*Mlh3*^{tm1Lpkn}/J heterozygotes were bred to maintain the colony and male homozygotes were used for Drop-seq experiments.
4. B6;129S7-*Hormad1*^{tm1Rajk}/Mmjax heterozygotes were bred to maintain the colony and male homozygous knockouts were used for Drop-seq experiments.
5. B6;129 *Cul4a*$^{-/-}$ mice were used for generating Drop-seq data

6. C57BL/6J CNP-EGFP BAC-TRAP mice were used for Drop-seq data

Single-cell Suspension Preparation

Mechanical Dissociation of Testes

Two different types of testicular dissociation protocols were used in this paper: enzymatic and mechanical. Both enzymatic and mechanical protocols were previously published in [182] and [71]. These methods were modified appropriately for single-cell RNA sequencing. For mechanical dissociation method, fresh testes were decapsulated in 1X DMEM and cut into small pieces (approximately 2-3mm^3). These tissue fragments were transferred to a 50µm Medicon, a tissue disaggregator and tissue fragments were dissociated in 1mL 1X DMEM for 5 minutes on Medimachine. The resulting single-cell suspension was aspirated from Medicon with a 3mL needless syringe. This dissociation/aspiration step was repeated three times and total of 3mL single-cell solution was retrieved. Then the single cells were filtered through sterile 40um strainers twice and triturated for 1 minute with a wide orifice disposable Pasteur pipet. Cells were spun down at 500xg for 10 minutes at 4°C and re-suspended in 2mL 1X DMEM. Finally, cells were filtered once more with sterile 50um filter, adjusted to 100 cells/µl concentration, and placed on ice until processed for Drop-seq. Single-cell RNA sequencing experiments were performed within ~30 minutes of testes collection for mechanical dissociation.

Enzymatic Dissociation of Testes

Solutions necessary for enzymatic dissociation were prepared fresh prior to testes collection and these solutions are as follows: 120U/mL collagenase type I in 1X DMEM; 50mg/mL trypsin in 1mM HCl; 1mg/mL DNase I in 50% glycerol. For enzymatic dissociation method, decapsulated fresh testes were collected in 15mL conical tubes, one testis per tube. Each testis was dissociated in 6 mL of collagenase type I solution and 10µl of DNAse I solution with horizontal agitation at

120rpm for 15 minutes at 37°C. Tubules were decanted for 1 minute vertically at room temperature and supernatant was discarded. Another 4mL of collagenase type I solution, 50µl of trypsin solution and 10µl of DNAse I solution were added to each tube and incubated with horizontal agitation at 120rpm for 15 minutes at 37°C. Testicular tubules were triturated with a plastic disposable Pasteur pipet with a wide orifice for 3 minutes. Another 30µl of Trypsin solution and 150µl of DNAse I solution were added and incubated for 10 minutes with horizontal agitation at 120rpm. Then 400µl Fetal Bovine Serum (FBS) was added to deactivate dissociation enzymes. Finally, collected single-cell suspension was passed through 40µm filter twice and stored on ice until processing for Drop-seq. These cells were processed within 1.5 hour of the testes collection.

For digesting *Pou5f1*-EGFP mice testes, we adapted a protocol described previously[183]. Briefly, testicular tubules/fragments were incubated in 200µg/mL trypsin solution for 15- 20 minutes with intermittent pipetting followed by 300µl FBS addition for inactivating trypsin. Single-cells suspension was filtered through 50µm filters twice and stored on ice until FACS.

Isolation of Germ Cell Populations by Flow Cytometry

Hoechst-FACS for spermatocytes and spermatids

For isolation of major germ cell populations, we adapted a Hoechst-FACS protocol and sequential gating strategies described in (Lima et al. 2017)[87]. Briefly, 10µl Hoechst and 2µl of propidium iodide (PI) were added to single-cell suspension obtained from one testis and incubated at room temperature for 20 minutes. Then single-cell suspension was filtered through a 50um cell strainer. Cells were sorted and analyzed using Beckman Coulter MoFlo Legacy cell sorter and Summit Cell sorting software. First, debris were excluded based on forward scatter (FSC) and side scatter (SSC) plot pattern. Single cells were gated by adjusting FSC and pulse width threshold. Dead cells were gated and removed based on PI intensity. A minimum of 500,000 events were observed before

proceeding to gating on different germ cell populations. Then, cell count histogram was plotted based on Hoechst blue fluorescence and observed three peaks, representing haploid (1C), diploid (2C), and tetraploid (4C) populations. Then Hoechst-blue and Hoechst-red fluorescence intensities were plotted to refine spermatocytes and spermatids populations.

Spermatogonia type A

For isolating spermatogonia type A cells from the *Pou5f1-EGFP* reporter mice, cells were analyzed and sorted with the same cell sorter and software described above section. Similar sequential gating strategies were followed. Debris were excluded, single cells were gated and dead cells were excluded. Then, GFP+ cells were gated on a plot of GFP vs FSC.

Single-cell RNA Sequencing Library Generation

Drop-seq Procedure

Drop-seq sequencing libraries were generated according to the protocol described previously (Macosko et al. 2015). Cells and beads were diluted to co-encapsulation occupancy of 0.05. Two bead lots were used for generating Drop-seq data (For more details, see **Supplementary File 1**). Individual droplets were broken by perfluorooctanol, followed by bead harvest and reverse transcription of hybridized mRNA. After Exonuclease I treatment, aliquots of 2000 beads were amplified for 14 PCR cycles (all necessary PCR reagents and conditions were identical to Macosko et al. 2015). PCR products were purified using 0.6x AMPure XP beads and cDNA from each experiment was quantified by Tapestation analysis. 600pg of cDNA was tagmented by Nextera XT with the custom primers, P5_TSO_Hybrid and Nextera 70X. The single-cell sequencing library from each batch was either pooled with another batch or sequenced separately on the Illumina HiSeq2500 at 1.4pM or MiSeq at 8pM, with custom priming (Read1CustSeqB Drop-seq primer).

Histological Methods

Collection and Processing of Testes

For histological studies, testes were collected in 4% paraformaldehyde (PFA), incubated overnight at 4°C and washed with 70% ethanol. For hematoxylin and eosin staining, testes were collected in modified Davidson fixative and after 24-hour incubation at room temperature, tissues were transferred to Bouin's solution for another 24-hour incubation at room temperature. Fixed testes were dehydrated through a series of graded ethanol baths and embedded in paraffin. Then 5µm sections were cut on clean glass slides.

Hematoxylin & Eosin (HE) Staining

Hematoxylin & Eosin staining was performed on each mouse line (Wildtype, *Mlh3-/-, Hormad1-/-, Cul4a-/-*, and *CNP-EGFP*) to assess overall morphology of testicular tissue. Slides were deparaffinized with xylene and rehydrated through a series of graded ethanol bath to PBS. Standard HE staining protocol was adapted from Belinda Dana (Department of Ophthalmology, Washington University in St. Louis) and followed with Hematoxylin 560 and 1% Alcoholic Eosin Y 515.

Immunofluorescence Staining

Prior to immunofluorescence staining, antigen retrieval was performed by boiling slides in citric acid buffer for 20 minutes, and tissue sections were blocked in blocking solution (0.5% Triton X-100 + 2% goat serum in 1X PBS) for an hour at room temperature. Primary antibodies were diluted to antibody-specific dilution (see Key Resources Table) and incubated overnight at 4°C in a humid chamber. Then, slides were incubated in secondary antibodies (1:300 dilution) at room temperature for 4 hours in a humid chamber. After the secondary antibody incubation, sections were stained with Hoechst (1:500 dilution), washed with 1X PBS and mounted with ProLong Diamond Antifade Mountant for visualization under confocal microscope.

Computational Methods

Preprocessing of Drop-seq Data

Paired-end sequencing reads were processed, filtered and aligned as described in [31]. The specific steps and tools for this process is further outlined in Drop-seq Computational Cookbook(http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf). STAR aligner was used to map the processed reads to mouse genome [184]. A STAR indexed genome was generated using mm10 mouse genome and GRCm38 gene annotation (release version 76) with default setting. Following the alignment, digital gene expression (DGE) matrices were generated for each experimental batch [31].

Quality Control for Drop-seq Data

Combined raw DGEs were processed through a series of quality control and normalization steps. Cells with fewer than 200 UMI counts or fewer than 50 genes expressed were removed. Cells were also removed if their total UMI count or number of genes expressed was more than 1 standard deviation below the mean for that experiment. A tSNE reduction of this dataset revealed an amorphous homogeneous group characterized by a low library size, high mitochondrial gene expression and often co-expressed genes from early and late meiosis suggesting poor quality and or doublet cells and so these were removed. Cells with a normalized mt-Rnr2 expression of greater than 2 were also removed. After these steps 20,322 cells and 28,893 genes remained.

Genes in the lower third of expression means were then removed and cells were normalized by square root transformation of total transcript counts per cell and genes were normalized to unit variance. All expression values were capped to maximum of 10. This results in a final matrix of 20,322 cells by 19,262 genes with a sparsity of 93.8% and a median UMI count of 1312 per cell.

K-means clustering and differential expression analysis

K-means clustering was performed on the t-SNE result of SDA run (the version that removed likely components that represent batch effects and technical artifacts) using 'kmeans' in R, testing different numbers for 'k' (**Figure 1- figure supplement 2**) with maximum iterations set to 10,000. We first settled with 'k=42 which slightly over-clustered the data (i.e. created more clusters than necessary) and then merged clusters that are transcriptionally indistinguishable. Briefly, a classification hierarchy tree that places transcriptionally similar clusters together was built using BuildClusterTree() function in Seurat(v2.3.0). To test for which clusters to be merged, the out-of-bag error (OOBE) method from a random forest classifier was used (implemented in Seurat via AssessNodes() and MergeNode() functions). The classification error was computed for left or right cells on each node of the tree and top 5 nodes with high OOBE were merged to finally produce 32 clusters in 'merged' t-SNE plot in **Figure 1- figure supplement 2.**Then, differentially expressed markers for all k-means clusters were identified using FindAllMarkers() function in Seurat with 'min.pct' parameter set to '0.25' where genes that are detected in a minimum fraction of 0.25 cells will be tested for differentially expressed genes. This differentially expressed genes list was used for assigning cell-types to each k-mean cluster and generating a list of potential novel cell-type specific markers by extracting top 10 differentially expressed genes for each cell-type and removing genes that were already annotated in the literature. A selected number of markers on this list was validated using immunofluorescence.

Somatic Cell Population Heterogeneity Analysis using Seurat

Seurat (v2.3.0) was used to subset, re-cluster and visualize somatic cell population data from joint wild-type and mutant dataset. After subsetting somatic cells from the original k-means result joint data **(clusters 1, 2, 3, 4, 5, 8 & 9 from Figure 1 - Figure Supplement 2A**), the percentage of

mitochondrial genes was re-calculated and then a linear transformation was applied (using ScaleData() function in Seurat) while regressing out unwanted source of variations (percentage of mitochondrial genes, number of transcripts, number of genes and batch). PCA was performed on the scaled data to reduce the dimensionality of the data. A number of statistically significant principle components (PCs) for clustering purpose was determined by plotting and examining the variability explained by each PC in decreasing order (using PCElbowPlot() function in Seurat). For clustering somatic cells, we used PC=18 as an input for K-nearest neighbor (KNN) graph-based algorithm implemented in Seurat (FindClusters()) along with resolution parameter set to '0.5.' We used t-SNE to visualize the data and clustering result. Differentially expressed genes (DEGs) were identified using Seurat's FindAllMarkers() function with 'min.pct' parameter set to '0.25' where genes that are detected in a minimum fraction of 0.25 cells are tested for DEGs. The DEGs list was used for performing gene ontology enrichment analysis to retrieve a functional profile for each somatic cluster. p-values were corrected using Benjamini-Hochberg.

Sparse Decomposition of Arrays (SDA)

SDA v1.1 [88,185]was then run on the post-QC final matrix with 50 components for 10,000 iterations to confirm convergence (although in practice the results are almost identical after just 1,000 iterations). The number of components was chosen such that there were typically between 1 and 5 single cell components across runs. Briefly, SDA decomposes a DGE into a number of components represented by two matrices. The columns vectors of the first matrix indicate how much a given component is active in each cell and the rows of the second matrix indicate which genes are active in a given component. SDA convergence was confirmed using the change in free energy, as well as the change in fraction of posterior inclusion probabilities (PIPs, probability that a gene loading is not equal to zero i.e. not in the spike) less than 0.5. The distribution of PIPs, cell scores, and

gene loadings were also assessed. SDA was also run four further times with different seeds as well as with different number of components to verify stability of the results. Those components with a single high loading in one cell (1, 4, 14, 18, 46) were removed to visualize relationships between the components. To visualize and quantify the biological relationships among cells, t-SNE (without initial PCA step) was run on a version of the component scores matrix with likely technical artifacts and batch components removed, using a "perplexity" parameter of 50, and 1000 iterations (Rtsne package [186–188]). Technical components were manually identified as meeting one or both of the following criteria: two batches of the same mouse line had opposite or very different cell scores (components 6, 12, 22, 28, 29, 41) and or if the highest loading genes were all or mostly ribosomal or pseudogenes (components 9, 25, 43). To assess uncertainty in the t-SNE embedding, t-SNE was also run multiple times with different seeds (**Figure 3 - figure supplement 5**). We also performed dimensionality reduction using UMAP and confirmed that it gave a pseudotime embedding consistent with tSNE [189](**Figure 3 - figure supplement 5**).

Note that SDA components have arbitrary sign and must be interpreted through the combination of gene and cell signs. Gene loadings and cell scores with concordant signs results in a positive expression contribution from a component, whereas discordant signs results in negative contribution.

To generate a pseudo-timeline we used a similar approach to that implemented in SCUBA [190]. We iteratively fit a principal curve through the t-SNE plot with increasing degrees of freedom from 4 to 9, using the curve from the previous run as the starting point using the princurve package in R [191]. Each cell was then assigned to its closest position on this curve, to define pseudotime for that cell. Somatic cells and the *Hormad1-/-*X-activated cells (component 38 score > 3) were excluded during pseudotime construction, but the *Hormad1-/-*X-activated cells were given pseudotimes

post-hoc. Somatic cells were defined by thresholding the cell scores of somatic components (if the absolute cell score of a given cell passed any of the following component thresholds 26, 11, 3, 32, 45, 24 >2; 37 >1.5; 40 >1; or mt-Rnr2 expression >3).

The temporal order of components was determined by using a weighted mean of the pseudotime values, where the weights are the cell scores of the component. In addition, only those cells with an absolute cell score of greater than 2 contribute to the mean. To calculate simulated haploid non sharing (**Figure 6**) cells with pseudotime > -10000 were randomly split into two groups. The predicted X expression was calculated as Original X * PseudoTime/10000 + e, where is a random normal error with mean 0 and s.d. of 3.

Computational analysis was performed using R [192]. Gene ontology enrichment analysis was performed on the top 250 genes from each component (from each side) using the enrichGO function from the clusterProfiler R package in which p-values are calculated based on the hypergeometric distribution and corrected for testing of multiple biological process GO terms using the Benjamini-Hochberg procedure [193]. Plots were created using the ggplot2 package and extensions ggrepel, ggforce, ggseqlogo, ggnewscale, ggrastr, RColorBrewer, viridis, and cowplot [194–201].In addition the following R packages were used: data.table, Matrix (for sparse large matrix computations), biomaRt (for gene identifiers), shiny and shinycssloaders (for creating the interactive web application), ComplexHeatmap, bigmemory (for creating a low-memory shiny app), and MASS (for kernel density estimation) [202–209].

Components were clustered by tSNE, using either the absolute gene loadings or cell scores matrix (tSNE perplexity=2). Component names were then assigned based on known maker genes from the literature and cross checked for consistency against the distribution of components in t-SNE

space. Components representing batch effects were identified by plotting cell scores by experimental batch and checking for biological subgroups with opposing cell scores.

We also ensured the KO cells were not unduly affecting the estimated components by separately performing an SDA analysis with only WT cells (normalized separately but with the same parameters). The same number of iterations, number of components, and random seed, were used. To account for rotations of the results we performed a procrustean rotation on the WT loadings matrix with the mixed loadings matrix as the target. Procrustes rotation was performed using the R package vegan [210,211]. We correlated the gene loadings of the Mixed WT & KO SDA analysis with the WT only analysis (after rotation) and found strong correspondence for those WT components which contained many cells **(Figure 3 - figure supplements 2-4**).

Validation of SDA Imputation

Imputed gene expression values (the posterior means of the SDA model) were computed as the matrix product of the cell scores and gene loadings matrix from SDA.

In order to formally quantify the accuracy of SDA imputation, we performed a cross validation study comparing the ability of SDA imputation to correctly predict single cell gene expression data in a held out sample. First, we randomly split the post-QC RNA-sequencing reads from the full dataset into two batches: with 20% probability a read is assigned to the test dataset, and with 80% probability it is assigned to the training dataset. Next we create seven predictors of gene expression levels for each cell, using the training dataset: “Unimputed” uses the training data directly (scaled by the total UMI counts for each cell), “Mean cell” uses the sum of training reads for each gene across all cells to predict ranks (i.e. every cell has the same prediction), the matrix factorisation approaches SDA, ICA, PCA and NNMF were run on the normalised training data (normalised as

described above) and imputed values calculated as the matrix product of cell scores and gene loadings, MAGIC values were computed using the Rmagic package.

To compare the accuracy of the three predictors for gene expression imputation, we evaluate an objective function for each predictor and each cell, which we call the "rank prediction accuracy curve" or RPAC. The RPAC for each predictor is created by rank ordering all genes in a single cell by the predicted level of expression of those genes, from high-to-low, after reversing normalizations (**Figure 5**). For each rank (abscissa), the ordinate is the cumulative fraction of test data reads for all genes up to that rank (i.e. all genes with higher predicted expression than the current rank). The RPAC is similar in spirit to a receiver operating characteristic (ROC) curve. The area under the curve (AUC) for each RPAC is informative about prediction accuracy; a completely random predictor is expected to produce an AUC of 0.5, while a method with some predictive utility will have an AUC > 0.5. This allows us to prefer predictions with a higher AUC, although we note that (unlike for a ROC) even given perfect imputation, the maximum possible expected AUC is < 1, because the test data is sparse and so shows considerable noise relative to the unknown truth.

In order to identify differences between SDA and NNMF (the most similar alternative method), for each gene we calculated imputed expression for both methods (not using single cell components from SDA), and calculated Pearson correlation between the two methods. We then looked for enrichment (by FET, p=0.05 after correction for multiple testing by Bonferroni) of the 500 least correlated genes in both SDA and NNMF components, finding 7 enriched SDA components (3P, 16N, 4N, 10P, 50N, 46N, 8P) and 0 NNMF components. We show example genes from 3P and 50N in **Figure 5 D and F**.

NNMF analysis was performed using the NNLM R package [210]with 50 components, and a stop criterion of $1x10^{-5}$. ICA analysis was performed with the fastICA R package [212]with 50 components. PCA was performed using the R package flashpcaR with 50 components and divisor and standardization set to "none" [213]. MAGIC was performed with default parameters using the R package Rmagic 1.5.0. [90].

Quantification and statistical analysis

Apolipoprotein E (ApoE) Immunofluorescence Signal Quantification

To quantify the frequency of ApoE protein signal in wildtype and mutant animals, we counted the total number of intact testicular tubules present on slides and the number of tubules with ApoE protein signal using a confocal microscope at 20x. A Fisher's exact test was used to test the hypothesis that the frequency of ApoE-positive tubules was the same in wildtype, *Mlh3-/-*and *Hormad1-/-*strains.

Data and software availability

Raw data and processed files for Drop-seq experiments are available under GEO accession number GEO: GSE113293

R markdown files that enable simulating main steps of the analysis are available upon reasonable request. Custom R code used is available at www.github.com/MyersGroup/testisAtlasand archived at DOI: 10.5281/zenodo.3233958.

SDA is available from https://jmarchini.org/sda/

3.6 Acknowledgments

We thank Abul Usmani for assistance with mouse husbandry and advice on *Pou5f1*:GFP reporter animals, Jeffrey Milbrandt and the WashU Genetics Department Single Cell Program for support, Liang Ma for providing *Cul4a* -/- mice, Joe Dougherty for providing *Cnp* mice, and Katinka Vigh-Conrad for assistance with figures. We also thank the Alvin J. Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, MO, for the use of the High Speed Cell Sorter Core, which provided cell sorting service. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant #P30 CA91842. This work was supported by National Institutes of Health Grants R01HD078641 and R01MH101810 to D.F.C., and Wellcome Trust grants 098387/Z/12/Z to S.M. and 109109/Z/15/Zto D.W.

3.7 Author Contributions

M.J. and J.R. performed all Drop-seq experiments; D.W., M.J., D.C. and S.M. analyzed the data. J.R., S.A. and M.J. performed histology. D.W., M.J., S.M. and D.C. wrote the paper with input from all authors. D.C., J.M. and S.M. supervised the project.

Chapter 4: Conclusion

4.1 Curation of other male infertility models

Our work in previous chapters illustrate the power of single-cell gene expression analysis in understanding both normal spermatogenesis and genetics of male infertility. We believe that presented single-cell isolation method and single-cell analysis framework can effectively characterize spermatogenic failure in male infertility mouse models, providing far superior resolution than the current method of describing gonadal defects. For our initial attempt, we only selected a small number of very well-studied male infertility models (*Mlh3-/-, Hormad1-/- and Cul4a-/-*) to test the applicability and power of our approach. Therefore, one of the future directions that would be interesting to pursue is to apply our developed framework to other male infertility or subfertility mouse models. This will help us to curate different types and mechanisms of spermatogenic failure in male infertility models. In addition, careful yet extensive curation of male infertility in mouse can help us to extract useful insights on understanding male infertility conditions in humans. As part of our effort to generate comprehensive anthology of possible gonadal defects, we generated approximately 8,000 single-cell gene expression data using Drop-seq from *Csmd1-/-* mice that displays subfertility phenotype.

4.2 Application of other single-cell technology in male infertility models

Following the introduction of droplet based single-cell RNA-sequencing methods, a number of new methods that can extrapolate additional information at single-cell level emerged.[44] Another future direction is to expand the depth of single-cell data generated from male infertility models using one of these emerging technologies. One of which is single-cell nuclei RNA sequencing (snRNA-seq) and it has a number of advantages over the single-cell RNA-sequencing.[214] First, it potentially reduces dissociation related biases as the range of nuclei

size is narrower than the size of whole cells.[215] Since the nuclei sizes will not differ much from different cell types, there will be less single-cell isolation or capturing biases based on the cell sizes. In addition, it is also compatible with frozen tissues which makes this method as a popular choice for studying clinical samples.[43] Finally, the snRNA-seq can estimate more immediate and relevant key transcribed genes activity that may be involved in the pathology of gonadal defects. However, these advantages come at a cost of capturing less genes in comparison to single-cell RNA-sequencing.[215]

Another recently developed single-cell sequencing method is called MATQ-seq. [216] MATQ-seq and other similar methods are scRNA-seq technologies that are capable of capturing RNA molecules with both poly A tails and non-poly A tailed RNAs.[44216] This will allow scientists to perform comprehensive investigation of both protein and non-coding gene expression dynamics at single-cell resolution. There have been numerous studies that these non-coding genes play integral part of spermatogenesis homeostasis[217–219] so it would be very exciting to apply MATQ or other similar technologies to study male infertility models.

4.3 Intersecting single cell transcriptome with exome in male infertility patients

Single-cell RNA-sequencing has already proven to provide deeper insights on the cellular dynamics and diversity of many different tissue types. To fully exploit the utility of high-throughput single-cell transcriptome data, a diagnostic framework that incorporates exome-sequencing data and single-cell expression data will be very useful for creating a more detailed molecular diagnostic report of patient-specific spermatogenic failure. Incorporating molecular information such as which variants are present in patients, estimation of the variants' deleterious impacts, and examination of how these variants are disturbing spermatogenic processes will

allow clinics to create a more accurate and personalized diagnosis report for males with infertile conditions.

Our collaborators at the University of Utah have already collected testicular biopsies from 3 healthy male controls, 2 Klinefelter syndrome patients, and 2 patients with idiopathic infertility condition. Using these testicular biopsies, they have also generated single-cell RNA-sequencing data using 10X Genomics. They have also collected peripheral blood samples from these patients and sequenced their exome. We will first need to develop a method that combines the exome sequencing data with the scRNA-seq for improving the identification of causal disease mutations, using the Population Sampling Probability (PSAP) approach recently developed by our lab (**Figure 12**). [220]

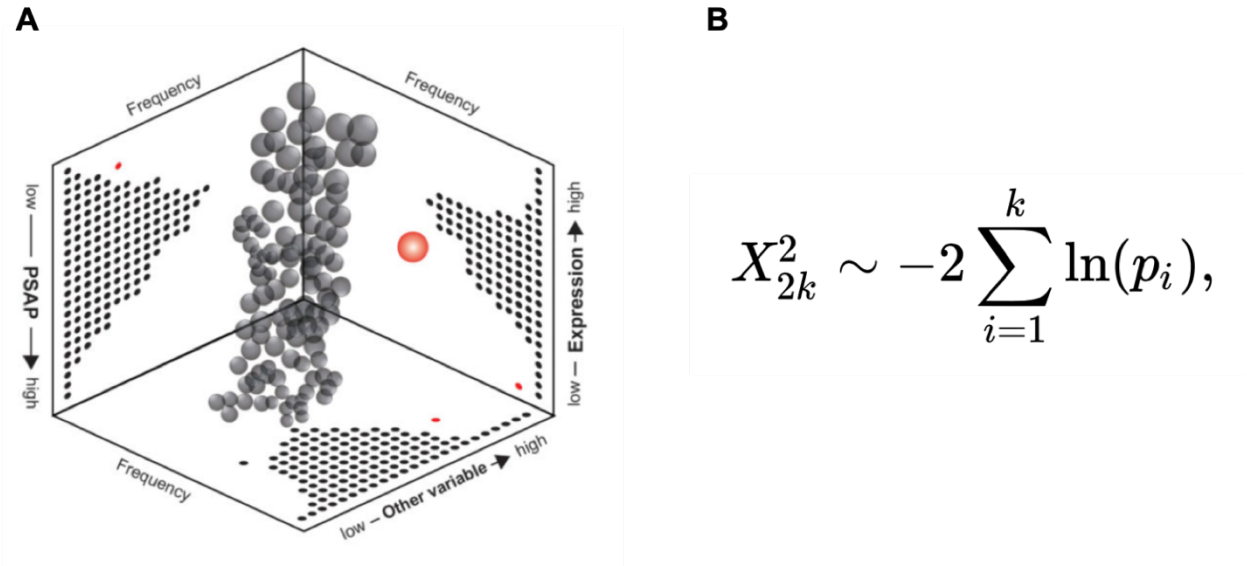


Figure 1: PSAP facilitates integrative analysis of single-cell transcriptome with exome from patients
(A) A joint analysis of a set of gene-based measurements (PSAP and RNA-seq p-values) for a single disease case (B) Fisher's method formula

Briefly, the original PSAP approach allows one to calculate well calibrated probabilities that a genotype was sampled from a "normal" population, by parameterizing null models of

genetic variation using large population databases such as ExAC. As PSAP is a gene-based method, it should be feasible to improve the identification of disrupted genes by combining exome-based PSAP p-values with other gene-based measurements from the affected individual. Here, we propose to construct PSAP p-values for two RNA-seq based measurements, RPKM and percent spliced in (pSI). Null distributions for these two statistics can be calculated from our own in-house testis RNA-seq. The RNA PSAP calculations can be made on a cell-type specific basis. We will combine the PSAP p-values from each cell type-specific RNA measurement with the exome PSAPs using Fisher's method to obtain a new summary PSAP for each gene (**Figure 12**). If p_i is the p-value from the i^{th} test (PSAP), then -2 times the log of the sum of these p-values follows a chi-square distribution under the null, with $2k$ degrees of freedom. With our proposed approach, we expect to delineate molecular defects on infertile patients, characterizing which cell types, pathways, and genes are affected in infertile patients and assessing which treatment may be most effective to individual patients.

4.4 Concluding remarks

We believe that our research will have significant impact on different areas of biology: (1) elucidating developmental biology of spermatogenesis, (2) characterizing how this tightly regulated process can go awry (2) demonstrating how single-cell analysis can expand our understanding of complex human disease and (3) laying the groundwork for how we effectively interpret existing defects of diseased state in comparison to healthy state. In the long term, we hope that our research will greatly advance our understanding of male reproductive biology and fill the gap between basic research science and translational medicine in male infertility

References

1. Dym, M. & Fawcett, D. W. Further observations on the numbers of spermatogonia, spermatocytes, and spermatids connected by intercellular bridges in the mammalian testis. *Biol. Reprod.* **4**, 195–215 (1971).
2. Braun, R. E., Behringer, R. R., Peschon, J. J., Brinster, R. L. & Palmiter, R. D. Genetically haploid spermatids are phenotypically diploid. *Nature* **337**, 373–376 (1989).
3. Cerilli, L. A., Kuang, W. & Rogers, D. A practical approach to testicular biopsy interpretation for male infertility. *Arch. Pathol. Lab. Med.* **134**, 1197–1204 (2010).
4. Rebourcet, D. *et al.* Sertoli cells maintain Leydig cell number and peritubular myoid cell activity in the adult mouse testis. *PLoS One* **9**, e105687 (2014).
5. Haider, S. G. Cell biology of Leydig cells in the testis. *Int. Rev. Cytol.* **233**, 181–241 (2004).
6. Maekawa, M., Kamimura, K. & Nagano, T. Peritubular myoid cells in the testis: their structure and function. *Arch. Histol. Cytol.* **59**, 1–13 (1996).
7. Soumillon, M. *et al.* Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports* **3**, 2179–2190 (2013).
8. Chen *et al.* ERM is required for transcriptional control of the spermatogonial stem cell niche. *Nature* **436**, 1030–1034 (2005).
9. Carelli, F. N. *et al.* The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**, 301–314 (2016).
10. Good, J. M., Giger, T., Dean, M. D. & Nachman, M. W. Widespread Over-Expression of the X Chromosome in Sterile F1 Hybrid Mice. *PLoS Genetics* **6**, e1001148 (2010).
11. Djureinovic, D. *et al.* The human testis-specific proteome defined by transcriptomics and antibody-

based profiling. *Mol. Hum. Reprod.* **20**, 476–488 (2014).

12. Guo, J. H., Huang, Q., Studholme, D. J., Wu, C. Q. & Zhao, Z. Transcriptomic analyses support the similarity of gene expression between brain and testis in human as well as mouse. *Cytogenet. Genome Res.* **111**, 107–109 (2005).
13. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
14. Schultz, N., Hamra, F. K. & Garbers, D. L. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12201–12206 (2003).
15. Zhou, Q. *et al.* Complete Meiosis from Embryonic Stem Cell-Derived Germ Cells In Vitro. *Cell Stem Cell* **18**, 330–340 (2016).
16. Yan, W. & McCarrey, J. R. Sex chromosome inactivation in the male. *Epigenetics* **4**, 452–456 (2009).
17. Hammoud, S. S. *et al.* Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460**, 473–478 (2009).
18. Peschon, J. J., Behringer, R. R., Brinster, R. L. & Palmiter, R. D. Spermatid-specific expression of protamine 1 in transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 5316–5319 (1987).
19. Diagnostic evaluation of the infertile male: a committee opinion. *Fertility and Sterility* **98**, 294–301 (2012).
20. Jungwirth, A. *et al.* European Association of Urology guidelines on Male Infertility: the 2012 update. *Eur. Urol.* **62**, 324–332 (2012).
21. O'Brien, K. L. O., O'Flynn O'Brien, K. L., Varghese, A. C. & Agarwal, A. The genetic causes of male factor infertility: A review. *Fertility and Sterility* **93**, 1–12 (2010).
22. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3**, 96ra77 (2011).

23. Carrell, D. T., Aston, K. I., Oliva, R., Emery, B. R. & De Jonge, C. J. The 'omics' of human male infertility: integrating big data in a systems biology approach. *Cell Tissue Res.* **363**, 295–312 (2016).

24. World Health Organization. *WHO Laboratory Manual for the Examination and Processing of Human Semen*. (2010).

25. Flannigan, R. & Schlegel, P. N. Genetic diagnostics of male infertility in clinical practice. *Best Pract. Res. Clin. Obstet. Gynaecol.* **44**, 26–37 (2017).

26. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, 63 (2016).

27. Chen, X. *et al.* Single-cell analysis at the threshold. *Nature Biotechnology* **34**, 1111–1118 (2016).

28. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610–620 (2015).

29. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

30. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

31. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

32. Chen, X., Teichmann, S. A. & Meyer, K. B. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. *Annual Review of Biomedical Data Science* **1**, 29–51 (2018).

33. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

34. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

35. Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci.*

Rep. **7**, 39921 (2017).

36. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours. doi:10.1101/165118
37. Buttner, M., Miao, Z., Wolf, A., Teichmann, S. A. & Theis, F. J. Assessment of batch-correction methods for scRNA-seq data with a new test metric. doi:10.1101/200345
38. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
39. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
40. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
41. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res.* **5**, (2016).
42. Wang, D. & Bodovitz, S. Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol.* **28**, 281–290 (2010).
43. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).
44. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **10**, 317 (2019).
45. Rodríguez-Casuriaga, R., Folle, G. A., Santiñaque, F., López-Carro, B. & Geisinger, A. Simple and efficient technique for the preparation of testicular cell suspensions. *J. Vis. Exp.* (2013). doi:10.3791/50102
46. Bryant, J. M., Meyer-Ficca, M. L., Dang, V. M., Berger, S. L. & Meyer, R. G. Separation of spermatogenic cell types using STA-PUT velocity sedimentation. *J. Vis. Exp.* (2013).

doi:10.3791/50648

47. Chang, Y.-F., Lee-Chang, J., Panneerdoss, S., Ii, J. M. & Rao, M. Isolation of Sertoli, Leydig, and spermatogenic cells from the mouse testis. *BioTechniques* **51**, (2011).
48. Getun, I. V., Torres, B. & Bois, P. R. J. Flow Cytometry Purification of Mouse Meiotic Cells. *Journal of Visualized Experiments* (2011). doi:10.3791/2602
49. Namekawa, S. H. *et al.* Postmeiotic Sex Chromatin in the Male Germline of Mice. *Current Biology* **16**, 660–667 (2006).
50. Yoshida, S. The first round of mouse spermatogenesis is a distinctive program that lacks the self-renewing spermatogonia stage. *Development* **133**, 1495–1505 (2006).
51. Laiho, A., Kotaja, N., Gyenesei, A. & Sironen, A. Transcriptome profiling of the murine testis during the first wave of spermatogenesis. *PLoS One* **8**, e61558 (2013).
52. Margolin, G., Khil, P. P., Kim, J., Bellani, M. A. & Camerini-Otero, R. D. Integrated transcriptome analysis of mouse spermatogenesis. *BMC Genomics* **15**, 39 (2014).
53. Bastos, H. *et al.* Flow cytometric characterization of viable meiotic and postmeiotic cells by Hoechst 33342 in mouse spermatogenesis. *Cytometry Part A* **65A**, 40–49 (2005).
54. Gaysinskaya, V. & Bortvin, A. Flow Cytometry of Murine Spermatocytes. *Current Protocols in Cytometry* 7.44.1–7.44.24 (2015). doi:10.1002/0471142956.cy0744s72
55. Gaysinskaya, V., Soh, I. Y., van der Heijden, G. W. & Bortvin, A. Optimized flow cytometry isolation of murine spermatocytes. *Cytometry Part A* **85**, 556–565 (2014).
56. Lassalle, B. `Side Population' cells in adult mouse testis express Bcrp1 gene and are enriched in spermatogonia and germinal stem cells. *Development* **131**, 479–487 (2003).
57. Watson, J. V., Nakeff, A., Chambers, S. H. & Smith, P. J. Flow cytometric fluorescence emission spectrum analysis of hoechst-33342-stained DNA in chicken thymocytes. *Cytometry* **6**, 310–315 (1985).
58. Sandhu, L. C., Warters, R. L. & Dethlefsen, L. A. Fluorescence studies of Hoechst 33342 with

supercoiled and relaxed plasmid pBR322 DNA. *Cytometry* **6**, 191–194 (1985).

59. Ellwart, J. W. & Dörmer, P. Vitality measurement using spectrum shift in Hoechst 33342 stained cells. *Cytometry* **11**, 239–243 (1990).
60. Smith, P. J., Nakeff, A. & Watson, J. V. Flow-cytometric detection of changes in the fluorescence emission spectrum of a vital DNA-specific dye in human tumour cells. *Experimental Cell Research* **159**, 37–46 (1985).
61. Steen, H. B. & Stokke, T. Fluorescence spectra of cells stained with a DNA-specific dye, measured by flow cytometry. *Cytometry* **7**, 104–106 (1986).
62. Falciatori, I. *et al.* Identification and enrichment of spermatogonial stem cells displaying side-population phenotype in immature mouse testis. *The FASEB Journal* **18**, 376–378 (2004).
63. Goodell, M. A. Isolation and functional properties of murine hematopoietic stem cells that are replicating in vivo. *Journal of Experimental Medicine* **183**, 1797–1806 (1996).
64. Liu, Y. *et al.* Fractionation of human spermatogenic cells using STA-PUT gravity sedimentation and their miRNA profiling. *Scientific Reports* **5**, (2015).
65. McCarrey, J. R. The epigenome--a family affair. *Science* **350**, 634–635 (2015).
66. Chowdhury, R., Bois, P. R. J., Feingold, E., Sherman, S. L. & Cheung, V. G. Genetic Analysis of Variation in Human Meiotic Recombination. *PLoS Genetics* **5**, e1000648 (2009).
67. Getun, I. V., Wu, Z. K., Khalil, A. M. & Bois, P. R. J. Nucleosome occupancy landscape and dynamics at mouse recombination hotspots. *EMBO reports* **11**, 555–560 (2010).
68. Roig, I. *et al.* Mouse TRIP13/PCH2 Is Required for Recombination and Normal Higher-Order Chromosome Structure during Meiosis. *PLoS Genetics* **6**, e1001062 (2010).
69. Rodríguez-Casuriaga, R. *et al.* Ultra-Fast and Optimized Method for the Preparation of Rodent Testicular Cells for Flow Cytometric Analysis. *Biological Procedures Online* **11**, 184–195 (2009).
70. Shimizu, Y. *et al.* A novel subpopulation lacking Oct4 expression in the testicular side population. *Int. J. Mol. Med.* **17**, 21–28 (2006).

71. Geisinger, A. & Rodríguez-Casuriaga, R. Flow cytometry for gene expression studies in Mammalian spermatogenesis. *Cytogenet. Genome Res.* **128**, 46–56 (2010).

72. Rodríguez-Casuriaga, R., Geisinger, A., Santiñaque, F. F., López-Carro, B. & Folle, G. A. High-purity flow sorting of early meiocytes based on DNA analysis of guinea pig spermatogenic cells. *Cytometry Part A* **79A**, 625–634 (2011).

73. Gonzalez, R. & Dobrinski, I. Beyond the Mouse Monopoly: Studying the Male Germ Line in Domestic Animal Models. *ILAR Journal* **56**, 83–98 (2015).

74. Castaneda, J. *et al.* Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice. *The EMBO Journal* **33**, 1999–2019 (2014).

75. Gan, H. *et al.* Integrative Proteomic and Transcriptomic Analyses Reveal Multiple Post-transcriptional Regulatory Mechanisms of Mouse Spermatogenesis. *Molecular & Cellular Proteomics* **12**, 1144–1157 (2013).

76. Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1839**, 155–168 (2014).

77. Schlatt, S. & Ehmcke, J. Regulation of spermatogenesis: An evolutionary biologist's perspective. *Seminars in Cell & Developmental Biology* **29**, 2–16 (2014).

78. Park, M. H. *et al.* Development of a high-yield technique to isolate spermatogonial stem cells from porcine testes. *Journal of Assisted Reproduction and Genetics* **31**, 983–991 (2014).

79. Reproductive Tissue Dissociation. Available at: http://www.worthington-biochem.com/tissuedissociation/Reproductive.html. (Accessed: 29th July 2016)

80. Hess, R. A. & de Franca, L. R. Spermatogenesis and Cycle of the Seminiferous Epithelium. in *Advances in Experimental Medicine and Biology* 1–15 (2009).

81. Dohle, G. R., Elzanaty, S. & van Casteren, N. J. Testicular biopsy: clinical practice and interpretation. *Asian J. Androl.* **14**, 88–93 (2012).

82. Chen, Y. *et al.* Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res.* **28**, 879–896 (2018).

83. Green, C. D. *et al.* A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. *Dev. Cell* **46**, 651–667.e10 (2018).

84. Lukassen, S., Bosch, E., Ekici, A. B. & Winterpacht, A. Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Sci. Rep.* **8**, 6521 (2018).

85. Ernst, C., Eling, N., Martinez-Jimenez, C. P., Marioni, J. C. & Odom, D. T. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. (2018). doi:10.1101/350868

86. Hermann, B. P. *et al.* The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. *Cell Rep.* **25**, 1650–1667.e8 (2018).

87. Lima, A. C. *et al.* A Standardized Approach for Multispecies Purification of Mammalian Male Germ Cells by Mechanical Tissue Dissociation and Flow Cytometry. *J. Vis. Exp.* (2017). doi:10.3791/55913

88. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48**, 1094–1100 (2016).

89. Marini, M. *et al.* Reappraising the microscopic anatomy of human testis: identification of telocyte networks in the peritubular and intertubular stromal space. *Sci. Rep.* **8**, 14780 (2018).

90. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716–729.e27 (2018).

91. Buaas, F. W. *et al.* Plzf is required in adult male germ cells for stem cell self-renewal. *Nat. Genet.* **36**, 647–652 (2004).

92. Goertz, M. J., Wu, Z., Gallardo, T. D., Hamra, F. K. & Castrillon, D. H. Foxo1 is required in mouse spermatogonial stem cells for their maintenance and the initiation of spermatogenesis. *J. Clin. Invest.* **121**, 3456–3466 (2011).

93. He, Z., Jiang, J., Hofmann, M.-C. & Dym, M. Gfra1 Silencing in Mouse Spermatogonial Stem Cells Results in Their Differentiation Via the Inactivation of RET Tyrosine Kinase1. *Biol. Reprod.* **77**, 723–733 (2007).

94. Kang, H. S. *et al.* Transcription Factor GLIS3: A New and Critical Regulator of Postnatal Stages of Mouse Spermatogenesis. *Stem Cells* **34**, 2772–2783 (2016).

95. Suzuki, H., Sada, A., Yoshida, S. & Saga, Y. The heterogeneity of spermatogonia is revealed by their topology and expression of marker proteins including the germ cell-specific proteins Nanos2 and Nanos3. *Dev. Biol.* **336**, 222–231 (2009).

96. Zheng, K., Wu, X., Kaestner, K. H. & Wang, P. J. The pluripotency factor LIN28 marks undifferentiated spermatogonia in mouse. *BMC Dev. Biol.* **9**, 38 (2009).

97. Oakberg, E. F. Duration of spermatogenesis in the mouse and timing of stages of the cycle of the seminiferous epithelium. *Am. J. Anat.* **99**, 507–516 (1956).

98. Boateng, K. A., Bellani, M. A., Gregoretti, I. V., Pratto, F. & Camerini-Otero, R. D. Homologous pairing preceding SPO11-mediated double-strand breaks in mice. *Dev. Cell* **24**, 196–205 (2013).

99. Ishiguro, K.-I. *et al.* Meiosis-specific cohesin mediates homolog recognition in mouse spermatocytes. *Genes Dev.* **28**, 594–607 (2014).

100. Keeney, S., Giroux, C. N. & Kleckner, N. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375–384 (1997).

101. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**, 876–879 (2010).

102. Parvanov, E. D., Petkov, P. M. & Paigen, K. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science* **327**, 835–835 (2010).

103. Baudat, F. *et al.* PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**, 836–840 (2010).

104. Zickler, D. & Kleckner, N. Recombination, Pairing, and Synapsis of Homologs during Meiosis. *Cold*

Spring Harb. Perspect. Biol. **7**, (2015).

105. Rankin, S. Complex elaboration: making sense of meiotic cohesin dynamics. *FEBS J.* **282**, 2426–2443 (2015).
106. Tu, Z. *et al.* Speedy A-Cdk2 binding mediates initial telomere-nuclear envelope attachment during meiotic prophase I independent of Cdk2 activation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 592–597 (2017).
107. Wang, Y. *et al.* The meiotic TERB1-TERB2-MAJIN complex tethers telomeres to the nuclear envelope. *Nat. Commun.* **10**, 564 (2019).
108. Ding, X. *et al.* SUN1 is required for telomere attachment to nuclear envelope and gametogenesis in mice. *Dev. Cell* **12**, 863–872 (2007).
109. Lukaszewicz, A., Lange, J., Keeney, S. & Jasin, M. Control of meiotic double-strand-break formation by ATM: local and global views. *Cell Cycle* **17**, 1155–1172 (2018).
110. Reinholdt, L. G. & Schimenti, J. C. Mei1 is epistatic to Dmc1 during mouse meiosis. *Chromosoma* **114**, 127–134 (2005).
111. Stanzione, M. *et al.* Meiotic DNA break formation requires the unsynapsed chromosome axis-binding protein IHO1 (CCDC36) in mice. *Nat. Cell Biol.* **18**, 1208–1220 (2016).
112. Robert, T. *et al.* The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. *Science* **351**, 943–949 (2016).
113. Vrielynck, N. *et al.* A DNA topoisomerase VI-like complex initiates meiotic recombination. *Science* **351**, 939–943 (2016).
114. Xu, Y., Greenberg, R. A., Schonbrunn, E. & Wang, P. J. Meiosis-specific proteins MEIOB and SPATA22 cooperatively associate with the single-stranded DNA-binding replication protein A complex and DNA double-strand breaks. *Biol. Reprod.* **96**, 1096–1104 (2017).
115. Ribeiro, J. *et al.* MEIOB and SPATA22 resemble RPA subunits and interact with the RPA complex to promote meiotic recombination. *Cell Biology* (2018).

116. Zhang, J., Fujiwara, Y., Yamamoto, S. & Shibuya, H. A meiosis-specific BRCA2 binding protein recruits recombinases to DNA double-strand breaks to ensure homologous recombination. *Nat. Commun.* **10**, 722 (2019).

117. Yang, F., Eckardt, S., Leu, N. A., McLaughlin, K. J. & Wang, P. J. Mouse TEX15 is essential for DNA double-strand break repair and chromosomal synapsis during male meiosis. *J. Cell Biol.* **180**, 673–679 (2008).

118. Martinez, J. S. *et al.* BRCA2 regulates DMC1-mediated recombination through the BRC repeats. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3515–3520 (2016).

119. Pacheco, S. *et al.* ATR is required to complete meiotic recombination in mice. *Nat. Commun.* **9**, 2622 (2018).

120. Widger, A. *et al.* ATR is a multifunctional regulator of male mouse meiosis. *Nat. Commun.* **9**, 2621 (2018).

121. Brown, M. S. & Bishop, D. K. DNA strand exchange and RecA homologs in meiosis. *Cold Spring Harb. Perspect. Biol.* **7**, a016659 (2014).

122. Brown, M. S., Grubb, J., Zhang, A., Rust, M. J. & Bishop, D. K. Small Rad51 and Dmc1 Complexes Often Co-occupy Both Ends of a Meiotic DNA Double Strand Break. *PLoS Genet.* **11**, e1005653 (2015).

123. Kovalenko, O. V., Wiese, C. & Schild, D. RAD51AP2, a novel vertebrate- and meiotic-specific protein, shares a conserved RAD51-interacting C-terminal domain with RAD51AP1/PIR51. *Nucleic Acids Res.* **34**, 5081–5092 (2006).

124. Dai, J., Voloshin, O., Potapova, S. & Daniel Camerini-Otero, R. Meiotic Knockdown and Complementation Reveals Essential Role of RAD51 in Mouse Spermatogenesis. *Cell Reports* **18**, 1383–1394 (2017).

125. Lee, K. Y. *et al.* MCM8-9 complex promotes resection of double-strand break ends by MRE11-RAD50-NBS1 complex. *Nature Communications* **6**, (2015).

126. Guiraldelli, M. F., Eyster, C., Wilkerson, J. L., Dresser, M. E. & Pezza, R. J. Mouse HFM1/Mer3 is required for crossover formation and complete synapsis of homologous chromosomes during meiosis. *PLoS Genet.* **9**, e1003383 (2013).

127. Guiraldelli, M. F. *et al.* SHOC1 is a ERCC4-(HhH)2-like protein, integral to the formation of crossover recombination intermediates during mammalian meiosis. *PLOS Genetics* **14**, e1007381 (2018).

128. Adelman, C. A. & Petrini, J. H. J. ZIP4H (TEX11) deficiency in the mouse impairs meiotic double strand break repair and the regulation of crossing over. *PLoS Genet.* **4**, e1000042 (2008).

129. Sun, X. *et al.* FancJ (Brip1) loss-of-function allele results in spermatogonial cell depletion during embryogenesis and altered processing of crossover sites during meiotic prophase I in mice. *Chromosoma* **125**, 237–252 (2016).

130. Rakshambikai, R., Srinivasan, N. & Nishant, K. T. Structural insights into Saccharomyces cerevisiae Msh4-Msh5 complex function using homology modeling. *PLoS One* **8**, e78753 (2013).

131. Syrjänen, J. L., Pellegrini, L. & Davies, O. R. A molecular model for the role of SYCP3 in meiotic chromosome organisation. *Elife* **3**, (2014).

132. Gómez-H, L. *et al.* C14ORF39/SIX6OS1 is a constituent of the synaptonemal complex and is essential for mouse fertility. *Nat. Commun.* **7**, 13298 (2016).

133. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, (2019).

134. Kong, A. *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nat. Genet.* **46**, 11–16 (2014).

135. Soh, Y. Q. S. *et al.* Meioc maintains an extended meiotic prophase I in mice. *PLoS Genet.* **13**, e1006704 (2017).

136. Abby, E. *et al.* Implementation of meiosis prophase I programme requires a conserved retinoid-independent stabilizer of meiotic transcripts. *Nat. Commun.* **7**, 10324 (2016).

137. Boekhout, M. *et al.* REC114 partner ANKRD31 controls number, timing and location of meiotic DNA breaks. (2018). doi:10.1101/425322

138. Papanikos, F. *et al.* ANKRD31 regulates spatiotemporal patterning of meiotic recombination initiation and ensures recombination between heterologous sex chromosomes in mice. (2018). doi:10.1101/423293

139. He, F. *et al.* Complex structure of the zf-CW domain and the H3K4me3 peptide. (2010). doi:10.2210/pdb2rr4/pdb

140. Rona, G. B., Eleutherio, E. C. A. & Pinheiro, A. S. PWWP domains and their modes of sensing DNA and histone methylated lysines. *Biophys. Rev.* **8**, 63–74 (2016).

141. Powers, N. R. *et al.* The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. *PLoS Genet.* **12**, e1006146 (2016).

142. da Cruz, I. *et al.* Transcriptome analysis of highly purified mouse spermatogenic cell populations: gene expression signatures switch from meiotic-to postmeiotic-related processes at pachytene stage. *BMC Genomics* **17**, 294 (2016).

143. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).

144. Turner, J. M. A. Meiotic sex chromosome inactivation. *Development* **134**, 1823–1831 (2007).

145. Turner, J. M. A. Meiotic Silencing in Mammals. *Annu. Rev. Genet.* **49**, 395–412 (2015).

146. Greenbaum, M. P., Iwamori, T., Buchold, G. M. & Matzuk, M. M. Germ cell intercellular bridges. *Cold Spring Harb. Perspect. Biol.* **3**, a005850 (2011).

147. Otto, S. P., Scott, M. F. & Immler, S. Evolution of haploid selection in predominantly diploid organisms. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15952–15957 (2015).

148. Veron, N. *et al.* Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder. *Genes Dev.* **23**, 2705–2710 (2009).

149. Martin-DeLeon, P. A. *et al.* 10.1186/1477-7827-3-32. *Reprod Biol Endocrinol* **3**, 32 (2005).

150. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).

151. Moretti, C., Vaiman, D., Tores, F. & Cocquet, J. Expression and epigenomic landscape of the sex chromosomes in mouse post-meiotic male germ cells. *Epigenetics Chromatin* **9**, 47 (2016).

152. Ito, C. & Toshimori, K. Acrosome markers of human sperm. *Anat. Sci. Int.* **91**, 128–142 (2016).

153. Sassone-Corsi, P. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science* **296**, 2176–2178 (2002).

154. Wang, W. *et al.* Proteomic analysis of murine testes lipid droplets. *Sci. Rep.* **5**, 12070 (2015).

155. Kuroda, N. *et al.* Distribution and role of CD34-positive stromal cells and myofibroblasts in human normal testicular stroma. *Histol. Histopathol.* **19**, 743–751 (2004).

156. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).

157. Oakberg, E. F. Duration of spermatogenesis in the mouse. *Nature* **180**, 1137–1138 (1957).

158. Wojtasz, L. *et al.* Mouse HORMAD1 and HORMAD2, two conserved meiotic chromosomal proteins, are depleted from synapsed chromosome axes with the help of TRIP13 AAA-ATPase. *PLoS Genet.* **5**, e1000702 (2009).

159. Shin, Y.-H. *et al.* Hormad1 mutation disrupts synaptonemal complex formation, recombination, and chromosome segregation in mammalian meiosis. *PLoS Genet.* **6**, e1001190 (2010).

160. Daniel, K. *et al.* Meiotic homologue alignment and its quality surveillance are controlled by mouse HORMAD1. *Nat. Cell Biol.* **13**, 599–610 (2011).

161. Royo, H. *et al.* Evidence that meiotic sex chromosome inactivation is essential for male fertility. *Curr. Biol.* **20**, 2117–2123 (2010).

162. Ortega, M. M. SURVEILLANCE MECHANISMS IN MAMMALIAN MEIOSIS. (Universitat Autònoma de Barcelona, 2016).

163. Dubiel, W., Dubiel, D., Wolf, D. A. & Naumann, M. Cullin 3-Based Ubiquitin Ligases as Master

Regulators of Mammalian Cell Differentiation. *Trends Biochem. Sci.* **43**, 95–107 (2018).

164. Yin, Y. *et al.* The E3 ubiquitin ligase Cullin 4A regulates meiotic progression in mouse spermatogenesis. *Dev. Biol.* **356**, 51–62 (2011).

165. Lipkin, S. M. *et al.* Meiotic arrest and aneuploidy in MLH3-deficient mice. *Nat. Genet.* **31**, 385–390 (2002).

166. Goluža, T. *et al.* Macrophages and Leydig cells in testicular biopsies of azoospermic men. *Biomed Res. Int.* **2014**, 828697 (2014).

167. Frungieri, M. *et al.* Number, distribution pattern, and identification of macrophages in the testes of infertile men. *Fertil. Steril.* **78**, 298–306 (2002).

168. Holstein, A. F. Spermatophagy in the seminiferous tubules and excurrent ducts of the testis in Rhesus monkey and in man. *Andrologia* **10**, 331–352 (1978).

169. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10101–10106 (2000).

170. Bernardo, J. M. *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*. (Oxford University Press, 2003).

171. Kim, P. M. & Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* **13**, 1706–1718 (2003).

172. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4164–4169 (2004).

173. Shao, C. & Höfer, T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* **33**, 235–242 (2017).

174. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015–1030.e16 (2018).

175. Welch, J. *et al.* Integrative inference of brain cell similarities and differences from single-cell genomics. *Neuroscience* (2018).

176. Zhu, X., Ching, T., Pan, X., Weissman, S. M. & Garmire, L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* **5**, e2888 (2017).

177. Duren, Z. *et al.* Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7723–7728 (2018).

178. Kotliar, D. *et al.* Identifying Gene Expression Programs of Cell-type Identity and Cellular Activity with Single-Cell RNA-Seq. *Bioinformatics* (2018).

179. Stein-O'Brien, G. L. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34**, 790–805 (2018).

180. Löffler-Wirth, H., Kalcher, M. & Binder, H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. *Bioinformatics* **31**, 3225–3227 (2015).

181. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).

182. Getun, I. V., Torres, B. & Bois, P. R. J. Flow cytometry purification of mouse meiotic cells. *J. Vis. Exp.* (2011). doi:10.3791/2602

183. Garcia, T. & Hofmann, M.-C. Isolation of undifferentiated and early differentiating type A spermatogonia from Pou5f1-GFP reporter mice. *Methods Mol. Biol.* **825**, 31–44 (2012).

184. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

185. Hore, V. Latent Variable Models for Analysing Multidimensional Gene Expression Data. (The University of Oxford, 2015).

186. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).

187. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

188. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. (2015).

189. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

190. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5643–50 (2014).

191. Hastie, T. & Stuetzle, W. Principal Curves. *J. Am. Stat. Assoc.* **84**, 502 (1989).

192. R Core Team. R: A Language and Environment for Statistical Computing. (2018).

193. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

194. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).

195. Campitelli, E. *ggnewscale: Multiple Fill and Color Scales in `ggplot2`*. (2019). doi:10.5281/zenodo.2543762

196. Pedersen, T. L. ggforce: Accelerating 'ggplot2'. (2016).

197. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).

198. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2018).

199. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2014).

200. Petukhov, V. ggrastr: Raster layers for ggplot2. (2018).

201. Garnier, S. viridis: Default Color Maps from 'matplotlib'. (2018).

202. Bates, D. & Maechler, M. Matrix: Sparse and Dense Matrix Classes and Methods. (2018).

203. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

204. Kane, M. J., Emerson, J. & Weston, S. Scalable Strategies for Computing with Massive Data. *Journal of Statistical Software* **55**, 1–19 (2013).

205. Sali, A. shinycssloaders: Add CSS Loading Animations to 'shiny' Outputs. (2017).

206. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S. (2002).

207. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. & McPherson, J. shiny: Web Application Framework for R. (2018).

208. Dowle, M. & Srinivasan, A. data.table: Extension of `data.frame`. (2019).

209. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

210. Lin, X. & Boutros, P. C. *NNLM: Fast and Versatile Non-Negative Matrix Factorization*. (2019).

211. Oksanen, J. *et al. vegan: Community Ecology Package*. (2019).

212. Marchini, J. L., Heaton, C. & Ripley, B. D. fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit. (2017).

213. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**, e93766 (2014).

214. Habib, N. *et al.* DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. doi:10.1101/115196

215. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, e0209648 (2018).

216. Sheng, K., Cao, W., Niu, Y., Deng, Q. & Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature Methods* **14**, 267–270 (2017).

217. Luk, A. C.-S., Chan, W.-Y., Rennert, O. M. & Lee, T.-L. Long noncoding RNAs in spermatogenesis: insights from recent high-throughput transcriptome studies. *Reproduction* **147**, R131–41 (2014).

218. Liu, K., Mao, X., Chen, Y., Li, T. & Ton, H. Regulatory role of long non-coding RNAs during reproductive disease. *Am. J. Transl. Res.* **10**, 1–12 (2018).

219. Holt, J. E., Stanger, S. J., Nixon, B. & McLaughlin, E. A. Non-coding RNA in Spermatogenesis and Epididymal Maturation. *Adv. Exp. Med. Biol.* **886**, 95–120 (2016).

220. Wilfert, A. B. *et al.* Genome-wide significance testing of variation from single case exomes. *Nat.*

Genet. **48**, 1455–1461 (2016).